CHAPTER 12

PRINCIPAL COMPONENTS ANALYSIS AND RECEPTOR MODELS IN ENVIRONMENTAL FORENSICS

Glenn W. Johnson, Robert Ehrlich, and William Full

12.1 INTRODUCTION

The identification of chemical contaminant sources is a common problem in environmental forensic investigations. Successful inference of sources depends on sampling plan design, sample collection procedures, chemical analysis methods, and knowledge of historical industrial processes in the study area. However, in complex situations where multiple sources contribute similar types of contaminants, even careful project planning and design may not be enough. If sources cannot be linked to a unique chemical species (i.e., a tracer chemical), then mapping the distributions of individual contaminant concentrations is insufficient to infer source. If, however, a source exhibits a characteristic 'chemical fingerprint' defined by diagnostic proportions of a large number of analytes, source inference may be accomplished through analysis of multiple variables; that is, through use of multivariate statistical methods. The objective of a multivariate approach to chemical fingerprinting is to determine (1) the number of fingerprints present in the system, (2) the multivariate chemical composition of each fingerprint, and (3) the relative contribution of each fingerprint in each collected sample.

Development of numerical methods to determine these parameters has been a major goal in environmental chemometrics and receptor modeling for more than 20 years. The result has been development of a series of procedures designed to accomplish this. As we shall see, procedures developed early in this history are most useful in solving relatively simple problems. Later procedures are designed to solve more general problems, which take into account complications such as bad data, commingled plumes (i.e., mixing of source fingerprints), and the presence of sources not assumed or anticipated at the start of an investigation. The objective of this chapter is to discuss this family of procedures in terms of their strengths and limitations, and in order to guide the working environmental scientist in the use of appropriate procedures.

12.1.1 PHILOSOPHY AND APPROACH: A CASE FOR EXPLORATORY DATA ANALYSIS

In terms of experimental design, the source apportionment problem in environmental forensic investigations falls between two extremes. At one extreme, all potential sources are known in terms of their chemical composition, location, history, and duration of activity. At the other extreme, none of these are known with any certainty. Chemicals at the receptor (e.g., estuary sediments, groundwater at a supply well) may be the result of activities long absent from the vicinity of the site.

In the first case (*a priori* knowledge of all sources) the problem is a relatively simple one. Appropriate sampling locations can be determined using a conventional experimental design, which is part of conventional experimental statistics. Determination of contribution of each source can be extracted using a variety of linear methods, such as chemical mass balance receptor models (see Chapter 11 of this volume). However, even when the contributing sources are known, environmental forensic investigations often prove to be more complex than initially anticipated. Chemicals in the environment may not retain their original composition. That is, chemical compositions may change over time by processes such as biodegradation and volatilization. This is true even for relatively recalcitrant contaminants such as polychlorinated biphenyls and dioxins (Bedard and Quensen, 1995; Chiarenzelli *et al.*, 1997). The result of degradation will be resolution of one or more fingerprints, not originally anticipated.

At the other extreme, where nothing is known with certainty, potential sources may be suspected, but samples of the sources (i.e., fingerprint reference standards) may not have been collected, and may not exist in the literature. The industrial history of a region may be imperfectly known. Often, a small, low profile operation may be a major but completely overlooked source of contamination. For cases towards this end of the spectrum, we must take leave of the elegance of conventional experimental statistics, and move into the realm of exploratory data analysis (EDA). The fundamental difference between these two approaches (experimental statistics and EDA) is that former is associated with creation of explicit hypotheses, and evaluation of data in terms of well-defined tests and strong probabilistic arguments. In contrast, the objective of EDA is to find patterns, correlations and relationships in the data itself, with few assumptions or hypotheses (Tukey, 1977). If the fruits of an EDA result in a map where the concentrations of a multivariate fingerprint increase monotonically towards an effluent pipe, and the fingerprint composition is consistent with the process associated with that source, the obvious inference is that the potential source is the actual source. We recognize that we are not working in the realm of classical statistics or formal hypothesis testing, and that EDA is based on less rigorous probabilistic statements. However, such an approach should not be construed as 'second best'. In environmental forensics, an EDA approach may be the only valid option. The cost of highly rigorous probabilistic methods is often a set of narrow assumptions regarding the structure of the data. Such methods cannot be supported if those critical assumptions cannot be safely assumed. Moreover, in environmental forensic investigations, we often lack sufficient information to know what hypotheses to test. Thus, at least at the beginning of a project, an EDA approach is usually most prudent.

Regardless of the data analytical strategy chosen, another important consideration is the presence of bad or questionable data. Common problems with environmental chemical data include the following: (1) chemical analyses performed by different laboratories or by different methods, which may introduce a systematic bias; (2) the presence of data at concentrations at or below method detection limits; (3) the presence of coelution (non-target analytes that elute at the same time as a target analyte; and (4) the ever-present problem of error in data entry, data transcription, or peak integration.

Unfortunately such errors rarely manifest themselves as random noise. More often, they contribute strong systematic variability. If unrecognized, the result may be derivation of 'fingerprints' which have little to do with the true sources. Therefore, a necessary adjunct to any data analysis in environmental forensics is vigilant identification of outliers. As we proceed with a discussion of fingerprinting methods, a major consideration in that regard must be inclusion of vigilant outlier identification and data cleaning procedures. If such an effort results in deletion or modification of data, the data must be clearly identified, and justification for the action must be provided in the narrative that accompanies the analyses.

12.1.2 FORMAL DESCRIPTION OF THE RECEPTOR MODELING PROBLEM

Receptor modeling in environmental forensics involves the inference of sources and their contributions through analysis of chemical data from the ambient environment (Gordon, 1988; Hopke, 1991). The objectives are to determine (1) the number of chemical fingerprints in the system; (2) the chemical composition of each fingerprint; and (3) the contribution of each fingerprint in each sample. The starting point is a data-table of chemical measurements in samples collected from the receptor (e.g., estuarine sediments, ambient air in a residential area). These data are usually provided in spreadsheet form where rows represent samples and columns represent chemical analytes. To the multivariate data analyst this table is a matrix. We will refer to the original data table as the m row by n column matrix **X**, where m is the number of samples and n is the number of analytes. We wish to know the number of fingerprints present (k) and chemical composition of each (objectives 1 and 2 above). This can be expressed as a matrix **F**, which has k rows and n columns. We also wish to know a third matrix (**A**), which has m rows and k columns, and represents the contribution of each fingerprint in each sample (objective 3 above). Thus the following linear algebraic equation formally expresses the receptor modeling problem.

$$\text{Matrix dimensions} \qquad \underset{(m\times n)}{\mathbf{X}} = \underset{(m\times k)}{\mathbf{A}} \; \underset{(k\times n)}{\mathbf{F}} \tag{12.1}$$

Given our data table we have three knowns (**X**, *m* and *n*), and three unknowns: (1) the number of fingerprints (*k*); (2) the fingerprint compositions (**F**); and (3) the contributions of each fingerprint in each sample (**A**). In the rare case where fingerprints are known *a priori*, both *k* and **F** are known, and only one unknown (**A**) remains. If this is the case, source compositions are contained within a 'training data set' consisting of the compositions of all potential sources, and solving Equation 12.1 for **A** is straightforward. The problem can be solved by regression techniques, which are the basis of chemical mass balance (CMB) approaches (see Chapter 11). Unfortunately, in environmental forensics investigations, we seldom have the luxury of *a priori* knowledge of contributing sources. The use of a training data set constitutes, if not a formal hypothesis test, at least an implicit hypothesis. If at all possible, we would like to derive source patterns directly from analysis of ambient data. That is, we want to employ chemometric methods that are 'self-training'.

While the number of chemometric methods available to us is large, the nature of environmental forensics investigations dictates the use of methods with very special features: they must allow for the conceptual model of mixtures of multiple sources; and they should allow resolution of contributing chemical fingerprints without *a priori* assumption of the number, chemical composition or geographic/temporal distribution. Finally, the results must be interpretable in a scientific context. Several commonly used chemometric methods, which satisfy the above considerations will be presented below.

12.1.3 DEMONSTRATION DATA SETS

The methods demonstrated in this chapter are illustrated using polychlorinated biphenyls (PCB) data. PCBs are a group of chlorinated organic compounds which are commonly the focus of environmental forensic investigations. PCBs were widely used in commercial and industrial settings for much of the twentieth century. Commercial PCB products were marketed under the trade-name Aroclor by Monsanto (the former US manufacturer). Commercial applications of PCBs included their use in fluorescent light ballasts, carbonless copy paper, and as dielectric fluids in electrical transformers and capacitors. PCBs are of concern in environmental studies because they are persistent, toxic and tend to bioaccumulate in tissues of higher predators (Tanabe *et al.*, 1987). PCBs are used for demonstrations in this chapter because they are a group of contaminants that typically require a multivariate approach to data analysis. Commercial PCB formulations (e.g., Aroclors) have unique chemical compositions composed of multiple PCB congeners, but no single congener is a diagnostic tracer for a specific Aroclor.

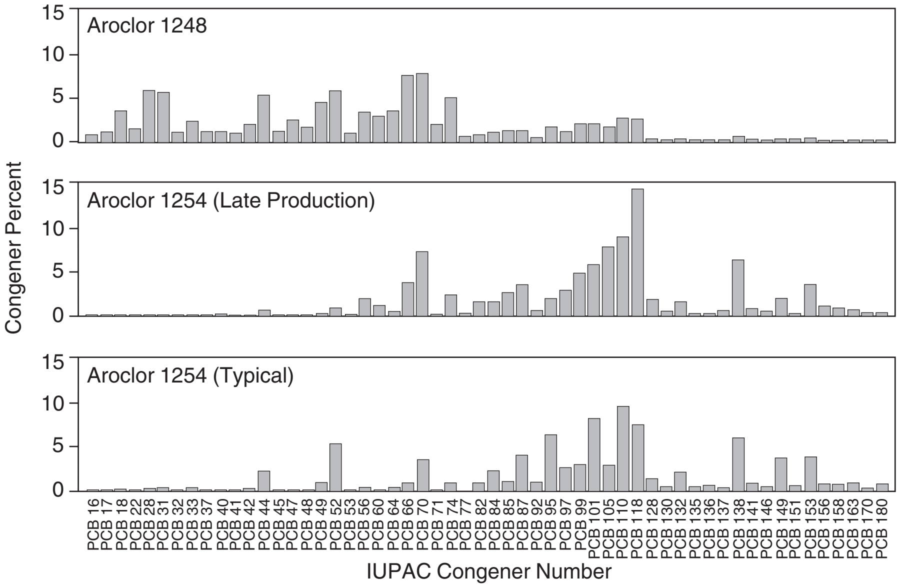
Two synthetic data sets will be used for demonstration purposes. Each of these data sets represents a three-source system. The congener patterns for the three-source fingerprints were taken from Aroclor standard compositions reported by Frame *et al.* (1996). The source compositions are Aroclor 1248 (Frame sample G3.5) and two variants of Aroclor 1254 (Frame samples A4 and G4). Frame *et al.* (1996) first reported markedly different congener patterns for different lots of Aroclor 1254. Subsequent investigations by Frame indicated that the atypical Aroclor 1254 was the result of a late production change in the Aroclor 1254 manufacturing process that occurred in the early 1970s (Frame, 1999). The congener compositions of these three-source fingerprints are shown as bar-graphs on Figure 12.1.

The two variants of Aroclor 1254 were used for this demonstration because it provides a typical example of the surprises often encountered in environmental forensic investigations. For example, given a situation where (1) the production history of an area was well established; (2) it was known with great certainty that only two PCB formulations were ever used at a site: Aroclor 1248 and Aroclor 1254; and (3) the data analyst was not aware of the two Aroclor 1254 variants; he/she might reasonably make an *a priori* assumption of a two-source system. Clearly, in this case, that assumption would be incorrect. In environmental forensics investigations, even the simplest, well-understood systems can yield surprises.

These three Aroclor compositions were used to create two synthetic data sets. The first of the two (Data Set 1) is quite simple, almost to the point of being unrealistic as an analogue to environmental forensics investigations.

Figure 12.1

Congener compositions of three PCB product formulations (Aroclors) used to create the artificial data sets used for demonstration in this chapter (data from Frame et al.*, 1996).*



However, it is instructive in that it provides a good intuitive understanding of principal components analysis (PCA), which is the mathematical basis of all the methods presented here. Data Set 1 is a 24-sample, 56-congener matrix. The data are simple in that each sample represents a contribution from one and only one source. That is, it is a strongly clustered data set, with no samples representing mixtures of two or of three Aroclors. To create some inter-sample variability within each of the three source categories, random Gaussian noise was added to each sample in the data set. Data Set 1 is shown on Table 12.1.

The second data set (Data Set 2) is considerably more complex, and is more representative of data encountered in environmental forensics investigations. Data Set 2 is shown in Table 12.3, and was created such that:

1 All samples are *mixtures* of the three Aroclor compositions. That is, no sample in the data set represents a 100% contribution from a single source. Varying contributions of multiple sources impacts every sample. The two matrices **A** and **F** (Equation 12.1) used to calculate the original noise free matrix are shown in Table 12.2. For ease of presentation, the transposed 56 × 3 matrix $\mathbf{F}^t$ (rather than the original matrix **F**) is shown in Table 12.2.
2 Ten percent Gaussian noise was added to simulate random error.
3 The data are represented in units of concentration (ng/g), and a method detection limit was established for each sample. As such, many low concentration matrix elements are 'censored' as a function of the detection limits. Non-detects are indicated in Table 12.3 with a 'U' qualifier. For subsequent numerical analyses, we adopt the common practice of replacing non-detect values with half the detection limit.
4 Data transcription errors were added to one sample (Sample 22).
5 For the congener PCB 141, a coelution problem was introduced in 35 of the 50 samples. Coelution of non-PCB peaks with PCB congeners during gas chromatographic analysis is a common problem in PCB chemistry. The coelution simulated here represents one such example. The pesticide *p,p*′-DDT elutes very close to PCB 141 on a Chrompack CP-SIL5-C18 GC column (R. Wagner, personal communication). Therefore, if *p,p*′-DDT is present in a sample undergoing congener specific PCB analysis, the *p,p*′-DDT could erroneously be reported as PCB congener IUPAC 141.

12.2 PRINCIPAL COMPONENTS ANALYSIS

12.2.1 PCA OVERVIEW

Principal components analysis (PCA) is a widely used method in environmental chemometrics, as it is in many scientific disciplines. PCA is used on its own, and as an intermediate step in receptor modeling methods. Before presenting the computational steps in PCA, it is useful to provide a more intuitive discussion, using Data Set 1 as an example. The objective of PCA is to reduce

Table 12.1

Data Set 1 (24 samples, 56 congeners) created by addition of Gaussian noise to three Aroclor compositions reported by Frame et al. *(1996). Units in percent.*

Sample	PCB 16	PCB 17	PCB 18	PCB 22	PCB 28	PCB 31	PCB 32	PCB 33	PCB 37	PCB 40	PCB 41	PCB 42	PCB 44
Aroclor 1248	0.76	0.75	2.70	1.16	6.79	4.98	1.05	1.50	1.04	0.74	0.64	1.90	6.34
Aroclor 1248	0.87	1.06	2.61	1.46	5.76	6.11	0.61	2.65	0.89	1.08	0.66	1.80	5.82
Aroclor 1248	0.66	0.99	3.72	1.65	6.46	6.68	0.97	2.64	1.22	0.85	0.63	1.95	5.89
Aroclor 1248	0.66	0.96	3.67	1.72	5.39	5.76	0.93	2.08	0.89	1.02	0.82	1.78	4.77
Aroclor 1248	0.65	1.01	3.95	2.03	6.17	3.78	1.14	2.46	1.05	0.92	0.87	1.85	6.89
Aroclor 1248	0.58	1.18	3.84	1.56	5.05	4.63	0.98	2.26	0.78	1.07	0.88	2.16	5.13
Aroclor 1248	0.70	0.96	3.61	1.30	3.80	5.65	1.07	2.96	1.02	1.07	0.89	1.50	5.93
Aroclor 1248	0.84	1.03	2.72	1.42	4.89	6.62	0.95	2.19	1.20	1.06	0.82	1.99	5.22
Aroclor 1254 (late production)	0.02	0.02	0.09	0.02	0.06	0.11	0.01	0.06	0.01	0.10	0.02	0.08	0.75
Aroclor 1254 (late production)	0.02	0.03	0.09	0.02	0.07	0.14	0.01	0.07	0.01	0.17	0.02	0.06	0.99
Aroclor 1254 (late production)	0.02	0.02	0.08	0.01	0.07	0.11	0.01	0.05	0.01	0.16	0.02	0.10	0.75
Aroclor 1254 (late production)	0.02	0.02	0.09	0.03	0.07	0.16	0.01	0.05	0.01	0.17	0.03	0.12	0.86
Aroclor 1254 (late production)	0.02	0.02	0.07	0.02	0.08	0.10	0.01	0.04	0.01	0.17	0.02	0.08	0.82
Aroclor 1254 (late production)	0.02	0.03	0.08	0.02	0.10	0.14	0.01	0.06	0.01	0.19	0.02	0.10	0.72
Aroclor 1254 (late production)	0.02	0.02	0.04	0.02	0.07	0.14	0.01	0.04	0.01	0.19	0.03	0.14	0.83
Aroclor 1254 (late production)	0.02	0.02	0.09	0.03	0.06	0.15	0.01	0.05	0.01	0.16	0.02	0.06	0.59
Aroclor 1254 (typical)	0.07	0.10	0.24	0.04	0.20	0.32	0.07	0.20	0.07	0.14	0.01	0.14	2.64
Aroclor 1254 (typical)	0.11	0.09	0.30	0.04	0.21	0.30	0.05	0.18	0.07	0.14	0.01	0.14	1.94
Aroclor 1254 (typical)	0.11	0.08	0.21	0.03	0.14	0.22	0.05	0.16	0.08	0.10	0.01	0.16	2.35
Aroclor 1254 (typical)	0.08	0.09	0.27	0.05	0.22	0.31	0.03	0.11	0.08	0.13	0.01	0.24	2.29
Aroclor 1254 (typical)	0.10	0.09	0.35	0.05	0.16	0.33	0.04	0.15	0.08	0.12	0.01	0.12	2.82
Aroclor 1254 (typical)	0.08	0.07	0.23	0.04	0.17	0.30	0.04	0.20	0.08	0.09	0.01	0.13	3.02
Aroclor 1254 (typical)	0.11	0.08	0.33	0.04	0.13	0.24	0.05	0.17	0.07	0.11	0.01	0.14	2.72
Aroclor 1254 (typical)	0.08	0.08	0.31	0.04	0.19	0.27	0.05	0.21	0.08	0.15	0.01	0.18	3.00

PCB 45	PCB 47	PCB 48	PCB 49	PCB 52	PCB 53	PCB 56	PCB 60	PCB 64	PCB 66	PCB 70	PCB 71	PCB 74	PCB 77	PCB 82
0.83	2.83	1.70	4.28	5.45	0.73	2.27	3.76	3.52	6.81	9.70	2.31	5.66	0.70	0.55
1.08	2.70	1.23	4.30	6.28	0.77	2.88	2.49	3.28	9.21	9.01	2.04	4.60	0.49	0.70
1.19	1.87	1.87	4.35	5.01	0.88	3.50	2.41	3.50	7.20	7.40	2.15	4.43	0.57	0.70
1.09	3.14	1.93	4.50	5.08	1.16	3.73	2.71	3.24	6.84	7.48	2.76	5.41	0.66	0.55
1.21	2.80	1.50	4.37	6.43	0.79	4.03	3.16	3.18	7.54	5.85	2.02	4.68	0.67	0.55
0.97	2.80	1.44	4.52	5.77	1.02	2.84	3.33	3.41	8.81	7.60	2.06	4.84	0.61	0.77
0.82	2.43	1.37	4.57	5.54	1.14	3.81	2.72	4.08	7.71	7.67	2.18	5.17	0.46	0.77
1.10	1.77	1.57	4.25	5.71	0.69	4.27	2.80	3.46	8.06	8.19	1.98	4.29	0.51	0.77
0.02	0.06	0.05	0.26	0.69	0.04	1.62	0.95	0.36	3.58	7.92	0.12	2.08	0.23	1.52
0.03	0.10	0.06	0.29	0.83	0.04	1.37	1.06	0.23	4.36	5.63	0.12	2.11	0.23	1.92
0.02	0.06	0.05	0.27	0.78	0.04	1.42	0.62	0.39	3.98	7.45	0.10	2.68	0.23	1.74
0.02	0.09	0.06	0.29	0.93	0.04	2.13	0.96	0.40	3.83	5.65	0.08	2.08	0.24	1.45
0.02	0.07	0.06	0.28	0.96	0.05	1.81	1.18	0.37	4.71	8.06	0.11	2.54	0.21	1.53
0.02	0.07	0.06	0.29	0.85	0.04	2.42	0.96	0.38	3.50	8.58	0.09	2.16	0.26	1.59
0.02	0.08	0.04	0.29	0.83	0.06	1.59	1.12	0.33	3.90	7.38	0.13	1.97	0.23	1.98
0.02	0.08	0.06	0.29	0.74	0.03	1.77	1.06	0.45	3.70	9.32	0.12	2.47	0.16	1.90
0.06	0.16	0.10	1.18	4.12	0.15	0.57	0.19	0.49	1.16	3.69	0.16	0.97	0.04	1.22
0.05	0.17	0.18	1.27	5.64	0.15	0.58	0.17	0.88	1.25	4.16	0.13	0.66	0.04	1.28
0.04	0.15	0.13	1.11	7.11	0.11	0.54	0.18	0.67	0.78	3.69	0.17	0.79	0.02	1.33
0.07	0.15	0.16	1.18	5.13	0.15	0.57	0.20	0.65	0.91	3.60	0.19	0.88	0.03	1.48
0.05	0.12	0.15	1.16	5.71	0.10	0.69	0.23	0.54	1.06	4.37	0.17	0.86	0.03	1.05
0.06	0.12	0.08	1.16	5.70	0.11	0.62	0.24	0.53	1.30	3.15	0.18	1.17	0.03	0.84
0.06	0.15	0.15	1.23	6.77	0.14	0.51	0.20	0.63	0.89	2.96	0.18	0.83	0.03	1.54
0.05	0.16	0.11	1.21	4.39	0.10	0.62	0.16	0.74	1.20	3.38	0.15	0.99	0.03	1.31

Table 12.1
Continued

Sample	PCB 84	PCB 85	PCB 87	PCB 92	PCB 95	PCB 97	PCB 99	PCB 101	PCB 105	PCB 110	PCB 118	PCB 128	PCB 130
Aroclor 1248	0.97	1.05	0.67	0.28	1.55	1.01	1.71	2.07	1.59	3.00	2.41	0.09	0.01
Aroclor 1248	0.98	1.17	0.95	0.25	1.36	1.12	2.00	2.26	1.44	1.87	1.94	0.08	0.01
Aroclor 1248	0.89	0.89	1.14	0.27	1.72	0.98	1.67	2.34	1.48	2.60	2.50	0.08	0.01
Aroclor 1248	1.00	1.15	1.13	0.21	1.57	1.20	1.96	2.43	1.57	2.65	2.00	0.06	0.01
Aroclor 1248	1.04	1.25	1.34	0.28	1.56	1.00	1.79	1.87	1.35	2.82	2.02	0.09	0.01
Aroclor 1248	1.05	1.10	1.43	0.23	1.38	1.10	1.93	1.44	1.77	2.51	3.01	0.09	0.01
Aroclor 1248	0.97	1.29	0.75	0.29	1.78	1.23	2.28	1.73	1.07	2.32	2.98	0.08	0.01
Aroclor 1248	0.80	1.27	1.48	0.23	1.75	0.94	1.88	1.75	1.36	3.14	2.68	0.08	0.01
Aroclor 1254 (late production)	1.83	2.27	3.60	0.63	1.99	2.78	4.81	6.55	8.02	11.26	14.48	1.39	0.50
Aroclor 1254 (late production)	2.02	2.50	3.03	0.62	2.02	2.39	5.32	6.78	9.52	8.60	13.85	2.09	0.64
Aroclor 1254 (late production)	1.67	2.66	3.82	0.62	1.59	2.89	4.53	6.59	7.18	9.10	17.61	1.94	0.34
Aroclor 1254 (late production)	1.72	2.69	3.92	0.64	2.18	2.78	4.30	7.52	5.89	7.90	14.42	2.56	0.65
Aroclor 1254 (late production)	1.47	2.22	2.85	0.64	1.92	3.80	3.68	4.33	7.12	10.01	14.88	2.14	0.50
Aroclor 1254 (late production)	1.98	2.74	3.90	0.45	1.80	2.38	5.50	6.91	6.24	9.93	10.95	2.09	0.52
Aroclor 1254 (late production)	1.95	2.31	4.00	0.48	2.02	2.44	5.43	4.85	7.26	7.78	16.30	1.99	0.55
Aroclor 1254 (late production)	1.39	3.04	3.03	0.57	1.55	3.43	5.26	5.67	7.24	10.00	11.55	2.09	0.60
Aroclor 1254 (typical)	1.98	1.71	4.30	1.27	7.23	2.34	3.34	4.98	3.28	10.74	9.63	1.70	0.66
Aroclor 1254 (typical)	2.15	1.33	4.78	1.75	5.74	3.00	4.40	7.80	3.21	9.91	8.14	1.59	0.51
Aroclor 1254 (typical)	2.74	1.36	4.17	0.88	7.27	3.16	3.51	7.94	2.59	10.14	10.10	1.22	0.48
Aroclor 1254 (typical)	2.99	1.23	4.74	1.19	6.53	3.10	3.42	7.67	2.70	10.28	7.99	1.67	0.69
Aroclor 1254 (typical)	2.22	1.49	3.93	1.17	6.10	2.45	2.64	9.72	3.65	9.61	10.14	1.46	0.75
Aroclor 1254 (typical)	1.82	1.89	3.96	1.41	7.39	2.75	3.81	7.08	3.54	12.26	7.92	1.48	0.62
Aroclor 1254 (typical)	2.91	1.77	4.63	1.02	6.99	2.03	3.25	8.63	2.65	9.86	8.89	1.98	0.68
Aroclor 1254 (typical)	2.02	1.66	4.23	1.02	6.43	2.48	3.54	7.81	3.62	12.39	6.52	1.95	0.58

PCB 132	PCB 135	PCB 136	PCB 137	PCB 138	PCB 141	PCB 146	PCB 149	PCB 151	PCB 153	PCB 156	PCB 158	PCB 163	PCB 170	PCB 180
0.13	0.05	0.06	0.02	0.42	0.09	0.05	0.33	0.08	0.43	0.05	0.04	0.09	0.06	0.24
0.16	0.04	0.05	0.01	0.46	0.09	0.05	0.39	0.08	0.39	0.04	0.04	0.11	0.07	0.17
0.13	0.05	0.06	0.02	0.34	0.10	0.05	0.40	0.08	0.36	0.05	0.04	0.10	0.10	0.21
0.13	0.05	0.05	0.02	0.55	0.10	0.05	0.36	0.10	0.44	0.05	0.04	0.08	0.09	0.25
0.14	0.04	0.07	0.02	0.44	0.10	0.05	0.27	0.07	0.36	0.04	0.04	0.11	0.10	0.19
0.17	0.04	0.05	0.03	0.38	0.08	0.04	0.26	0.06	0.45	0.02	0.04	0.07	0.06	0.29
0.16	0.04	0.06	0.02	0.40	0.11	0.06	0.34	0.08	0.49	0.04	0.05	0.08	0.10	0.29
0.13	0.05	0.06	0.02	0.46	0.11	0.05	0.30	0.10	0.49	0.03	0.04	0.09	0.09	0.23
1.10	0.31	0.19	0.60	6.32	0.59	0.52	1.93	0.26	3.24	1.30	1.25	0.81	0.33	0.30
1.72	0.35	0.28	0.53	7.08	0.78	0.49	2.31	0.29	2.65	0.93	1.36	0.78	0.46	0.53
1.36	0.28	0.29	0.61	5.92	0.64	0.46	1.90	0.31	3.11	0.84	0.91	0.84	0.31	0.43
2.09	0.39	0.32	0.63	7.98	0.74	0.43	2.40	0.26	4.00	1.20	0.93	0.70	0.53	0.38
1.57	0.20	0.26	0.51	6.74	0.74	0.56	2.04	0.22	3.95	1.04	1.35	0.95	0.38	0.48
1.72	0.25	0.23	0.76	6.08	0.58	0.46	2.21	0.26	4.99	1.35	1.31	0.80	0.39	0.47
2.02	0.34	0.30	0.68	6.48	0.63	0.52	1.83	0.21	4.25	1.10	0.94	0.88	0.37	0.55
2.07	0.24	0.30	0.61	6.35	0.63	0.64	2.07	0.28	4.02	1.19	0.94	0.90	0.46	0.42
2.87	0.59	0.69	0.34	6.85	1.22	0.56	4.79	0.81	5.44	1.00	0.87	1.11	0.55	0.67
3.12	0.64	0.96	0.47	5.83	1.35	0.69	3.55	0.87	3.37	0.90	0.88	1.38	0.65	0.84
1.75	0.66	0.79	0.37	6.29	1.16	0.77	3.46	0.76	3.98	1.06	0.91	0.86	0.42	0.65
2.35	0.69	0.74	0.46	7.57	1.30	0.67	3.65	0.82	4.08	0.90	0.64	1.27	0.63	0.77
2.20	0.63	0.78	0.44	6.10	0.92	0.47	3.24	0.75	4.42	0.62	0.96	1.21	0.40	0.79
2.84	0.65	0.85	0.40	4.56	1.17	0.81	3.11	0.64	4.91	1.08	0.96	1.20	0.45	0.68
2.39	0.61	0.61	0.50	5.50	0.80	0.68	3.26	0.62	4.82	0.82	0.76	1.36	0.62	0.89
2.54	0.61	0.72	0.48	6.33	1.22	0.59	4.43	0.75	4.20	0.91	1.23	1.14	0.66	0.72

Table 12.2

*Input matrices for artificial three-source PCB mixture. Multiplication by Equation 12.1 (**X** = **A*****F**) yields error free matrix **X**.*

Source Contributions Matrix (Mixing Proportions) [A]				End-Member Source Compositions Matrix [F^t] (Shown graphically in Figure 12.1)			
Sample Number	**Source 1 Aroclor 1248 (%)**	**Source 2 Late Production Aroclor 1254 (%)**	**Source 3 Typical Aroclor 1260 (%)**	**IUPAC Congener**	**Source 1 Aroclor 1248**	**Source 2 Late Production Aroclor 1254**	**Source 3 Typical Aroclor 1260**
Sample 1	3	43	54	1 PCB 16	0.75	0.02	0.10
Sample 2	22	26	51	2 PCB 17	0.98	0.02	0.09
Sample 3	59	13	28	3 PCB 18	3.46	0.09	0.27
Sample 4	71	11	18	4 PCB 22	1.45	0.02	0.04
Sample 5	3	38	59	5 PCB 28	5.86	0.06	0.21
Sample 6	59	24	17	6 PCB 31	5.76	0.12	0.30
Sample 7	45	7	48	7 PCB 32	0.98	0.01	0.05
Sample 8	41	2	57	8 PCB 33	2.33	0.05	0.17
Sample 9	19	37	44	9 PCB 37	1.00	0.01	0.08
Sample 10	60	35	5	10 PCB 40	0.97	0.16	0.13
Sample 11	20	74	5	11 PCB 41	0.79	0.02	0.01
Sample 12	55	12	33	12 PCB 42	1.88	0.10	0.16
Sample 13	48	8	44	13 PCB 44	5.36	0.72	2.50
Sample 14	17	72	11	14 PCB 45	0.96	0.02	0.05
Sample 15	44	13	43	15 PCB 47	2.54	0.08	0.15
Sample 16	44	36	20	16 PCB 48	1.62	0.05	0.13
Sample 17	48	50	2	17 PCB 49	4.39	0.28	1.19
Sample 18	15	57	29	18 PCB 52	5.87	0.89	5.81
Sample 19	47	47	7	19 PCB 53	0.93	0.04	0.13
Sample 20	14	1	85	20 PCB 56	3.36	1.83	0.59
Sample 21	40	8	52	21 PCB 60	2.81	1.02	0.19
Sample 22	46	45	10	22 PCB 64	3.50	0.39	0.64
Sample 23	50	22	28	23 PCB 66	7.60	3.84	1.09
Sample 24	49	45	6	24 PCB 70	7.78	7.36	3.77
Sample 25	63	27	10	25 PCB 71	1.96	0.12	0.16
Sample 26	3	67	30	26 PCB 74	4.92	2.36	0.91
Sample 27	24	26	50	27 PCB 77	0.55	0.22	0.03
Sample 28	57	7	36	28 PCB 82	0.65	1.65	1.20
Sample 29	24	12	65	29 PCB 84	0.96	1.70	2.51
Sample 30	67	6	27	30 PCB 85	1.20	2.68	1.38
Sample 31	50	21	29	31 PCB 87	1.17	3.68	4.31
Sample 32	7	79	13	32 PCB 92	0.26	0.61	1.39
Sample 33	7	46	47	33 PCB 95	1.51	1.98	6.75
Sample 34	3	79	18	34 PCB 97	1.02	3.00	2.83
Sample 35	18	38	44	35 PCB 99	1.91	4.88	3.26
Sample 36	38	28	34	36 PCB101	1.99	5.92	8.67
Sample 37	34	30	36	37 PCB105	1.53	7.95	3.23
Sample 38	51	20	29	38 PCB110	2.68	9.08	10.04
Sample 39	43	52	6	39 PCB118	2.47	14.65	7.94
Sample 40	38	0	62	40 PCB128	0.08	1.84	1.53
Sample 41	7	28	65	41 PCB130	0.01	0.54	0.65
Sample 42	16	4	80	42 PCB132	0.15	1.62	2.47
Sample 43	29	50	22	43 PCB135	0.04	0.30	0.66
Sample 44	21	51	28	44 PCB136	0.06	0.26	0.76
Sample 45	56	43	1	45 PCB137	0.02	0.56	0.45
Sample 46	24	42	34	46 PCB138	0.43	6.42	6.27

Table 12.2

Continued

Source Contributions Matrix (Mixing Proportions) [A]				End-Member Source Compositions Matrix $[F^t]$ (Shown graphically in Figure 12.1)			
Sample Number	**Source 1 Aroclor 1248 (%)**	**Source 2 Late Production Aroclor 1254 (%)**	**Source 3 Typical Aroclor 1260 (%)**	**IUPAC Congener**	**Source 1 Aroclor 1248**	**Source 2 Late Production Aroclor 1254**	**Source 3 Typical Aroclor 1260**
Sample 47	47	21	32	47 PCB141	0.09	0.74	1.06
Sample 48	71	6	23	48 PCB146	0.05	0.49	0.72
Sample 49	37	39	24	49 PCB149	0.35	1.96	3.94
Sample 50	68	23	8	50 PCB151	0.08	0.24	0.75
				51 PCB153	0.45	3.55	4.07
				52 PCB156	0.04	1.22	0.89
				53 PCB158	0.04	0.97	0.88
				54 PCB163	0.08	0.75	1.11
				55 PCB170	0.08	0.38	0.56
				56 PCB180	0.22	0.45	0.72

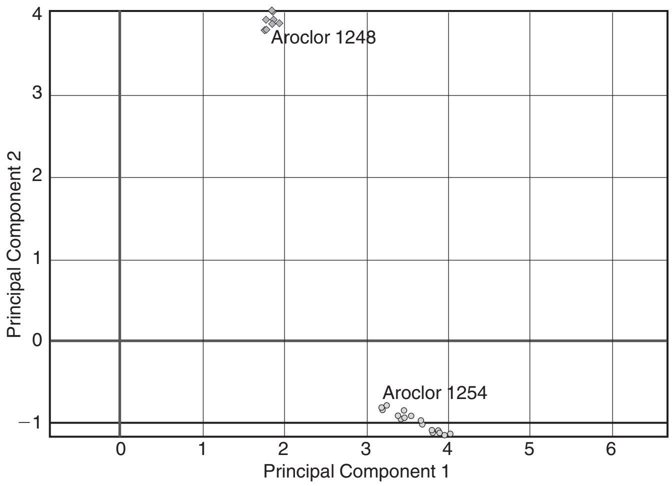


Figure 12.2

Two principal component scores plot of Data Set 1. Two PCs account for > 92% of the variance of Data Set 1, but are insufficient to allow distinction of the two Aroclor 1254 variants.

the dimensionality of a data set in which there are a large number of interrelated (i.e., correlated) variables. This reduction in dimension is achieved by transforming the data to a new set of uncorrelated reference variables (principal components or PCs). The PCs are sorted such that each in turn accounts for a progressively smaller percentage of variance within the data set. If nearly all variability between samples can be accounted for by a small number of PCs, then relationships between multivariate samples may be assessed by simple inspection of a two- or three-dimensional plot: a principal components scores plot. Figure 12.2 shows a two-PC scores plot for Data Set 1.

Table 12.3

Data Set 2 (50 Sample by 56 congeners) created by (1) multiplication of matrices **A** *and* **F** *(Table 12.2) as per Equation 12.1; (2) transfomation of to concentrations (ng/g); (3) addition of 10% random Gaussian noise; (4) censoring of the data based on a sample specific detection limit (censored data qualified as 'U'. Reported measurement is the detection limit); (5) simulation of a data transcription errors in Sample 22; and (6) simulation of DDT coelution with PCB 141 in a subset of samples.*

Sample	**PCB 16**	**Qualifier**	**PCB 17**	**Qualifier**	**PCB 18**	**Qualifier**	**PCB 22**	**Qualifier**	**PCB 28**	**Qualifier**	**PCB 31**	**Qualifier**	**PCB 32**	**Qualifier**	**PCB 33**	**Qualifier**
Sample 1	0.29		0.30		0.94		0.24	U	0.99		1.2		0.24		0.52	
Sample 2	0.66	U	0.66	U	1.6		0.66	U	2.8		2.8		0.66	U	1.3	
Sample 3	0.59		0.82		2.2		1.16		4.3		3.9		0.79		1.9	
Sample 4	1.5		1.7		6.9		2.70		12.2		9.6		1.6		4.5	
Sample 5	0.64		0.80		2.9		0.64	U	3.1		3.2		0.72		1.7	
Sample 6	1.2		1.7		5.4		2.2		9.2		9.8		1.6		3.2	
Sample 7	3.8		4.3		14.7		6.4		28.0		26.0		4.3		8.8	
Sample 8	2.7		2.6		10.2		3.2		15.4		17.9		2.9		8.2	
Sample 9	1.3	U	1.3	U	1.3	U	1.3	U	1.3	U	1.3	U	1.3	U	1.3	U
Sample 10	1.8		2.0		7.8		3.1		10.9		13.4		2.6		5.6	
Sample 11	1.4		1.3		5.8		2.4		8.6		7.9		1.6		3.4	
Sample 12	1.5		2.0		7.6		3.3		15.6		13.0		2.5		5.0	
Sample 13	2.4		3.2		11.0		4.4		15.7		18.9		3.2		7.5	
Sample 14	1.1		1.3		4.0		1.7		7.1		8.1		1.2		2.9	
Sample 15	1.5		2.2		5.2		2.8		11.3		12.1		2.1		4.9	
Sample 16	1.9		2.9		8.7		4.2		14.8		16.3		2.3		6.7	
Sample 17	1.1	U	1.1	U	1.4		1.1	U	2.3		2.6		1.1	U	1.1	U
Sample 18	0.60		0.76		3.1		1.1		4.2		4.4		0.65		2.02	
Sample 19	1.7		2.4		8.8		3.8		13.9		14.8		2.3		4.9	
Sample 20	3.5		3.6		12.8		3.9		15.9		15.6		2.9		6.9	
Sample 21	1.8		2.1		8.0		2.6		10.8		11.6		2.2		4.7	
Sample 22	3.0		3.7		17.4		5.8		20.3		20.8		4.2		9.0	
Sample 23	1.5		1.8		7.3		3.4		12.2		9.5		2.0		4.8	
Sample 24	2.0		2.3		8.1		3.4		15.2		14.3		2.7		5.0	
Sample 25	4.3		3.7		20.8		9.2		37.2		37.5		5.5		15.9	
Sample 26	0.50	U	0.50	U	1.6		0.50	U	2.0		2.3		0.50	U	1.2	
Sample 27	1.7		2.2		6.8		2.7		12.1		13.3		2.6		5.2	
Sample 28	1.6		2.0		7.2		2.2		11.2		11.0		2.0		4.2	
Sample 29	1.5		1.9		6.6		2.7		9.7		12.1		2.4		4.4	
Sample 30	1.8		2.3		7.5		3.2		14.8		12.7		2.3		5.5	
Sample 31	0.71		0.99		2.14		0.90		4.3		4.6		0.70		2.0	
Sample 32	1.2		1.4		5.2		2.2		8.0		9.3		1.5		3.9	
Sample 33	0.78	U	0.78	U	0.86		0.78	U	1.2		1.2		0.78	U	0.78	U
Sample 34	0.66	U	0.66	U	2.1		0.66	U	2.4		3.0		0.66	U	1.3	
Sample 35	1.8		1.8		8.0		2.8		11.3		10.7		2.1		5.2	
Sample 36	1.3		1.8		6.2		2.4		11.0		7.6		1.7		4.5	
Sample 37	1.0		1.6		4.0		2.4		6.6		8.6		1.5		2.9	
Sample 38	0.78		1.0		2.8		1.3		5.5		6.0		0.76		2.3	
Sample 39	0.72	U	1.1		3.0		1.4		5.5		4.6		0.92		1.8	
Sample 40	1.9		2.5		9.3		3.9		13.1		13.4		1.9		6.7	
Sample 41	0.46		0.44		1.7		0.52		2.2		3.0		0.41		1.1	
Sample 42	0.36	U	0.36	U	0.83		0.36	U	1.2		0.85		0.36	U	0.6	
Sample 43	0.74		0.83		3.3		1.2		6.0		5.1		0.85		2.6	
Sample 44	1.09		1.2		5.3		1.8		6.8		6.3		1.4		2.3	
Sample 45	2.6		2.2		10.3		3.8		20.2		19.2		3.0		8.1	
Sample 46	2.2		2.5		8.4		3.0		12.0		16.5		2.3		4.6	
Sample 47	2.3		2.4		10.0		4.2		13.3		12.5		2.7		6.9	
Sample 48	7.4		9.8		40.3		14.2		60.7		52.6		10.0		23.9	
Sample 49	2.5		3.5		14.1		5.4		21.0		19.7		4.0		8.3	
Sample 50	3.5		4.2		12.9		4.2		17.9		24.0		4.0		9.3	

PCB 37	Qualifier	PCB 40	Qualifier	PCB 41	Qualifier	PCB 42	Qualifier	PCB 44	Qualifier	PCB 45	Qualifier	PCB 47	Qualifier	PCB 48	Qualifier	PCB 49	Qualifier	PCB 52	Qualifier	PCB 53	Qualifier
0.30		0.63		0.24	U	0.66		6.0		0.24	U	0.69		0.50		3.6		12.4		0.41	
0.66	U	0.66	U	0.66	U	0.92		4.0		0.66	U	1.1		0.73		3.0		7.7		0.66	U
0.83		0.74		0.45		1.7		5.2		0.84		2.0		1.3		3.6		6.5		0.82	
1.9		1.8		1.4		3.3		10.8		1.9		4.6		3.8		9.4		14.6		2.0	
0.71		1.5		0.64	U	1.7		16.8		0.64	U	1.9		1.6		9.0		34.6		1.1	
1.6		1.6		1.2		3.1		8.6		1.6		4.6		3.1		7.8		14.2		1.3	
4.4		5.2		3.3		8.7		35.5		4.2		11.7		5.1		23.1		47.7		3.8	
3.0		3.1		2.2		6.1		27.1		2.6		7.5		4.4		14.2		39.2		2.86	
1.3	U	1.3	U	1.3	U	1.3	U	2.2		1.3	U	1.3	U	1.3	U	1.4		3.9		1.3	U
2.2		1.8		1.6		3.8		14.2		1.8		5.9		3.5		9.3		14.7		2.3	
1.4		2.5		1.2		3.2		12.8		1.4		3.9		2.9		8.3		14.9		1.8	
2.4		2.5		1.8		5.2		14.7		2.4		6.0		4.3		10.7		18.2		2.6	
2.9		3.0		2.6		5.0		24.5		3.0		8.1		3.7		16.6		32.2		2.4	
1.4		2.1		1.1		2.8		10.7		1.1		3.4		2.6		6.9		13.7		1.4	
1.8		1.8		1.5		4.5		13.8		1.7		4.9		3.3		10.7		25.3		1.7	
2.5		2.8		1.9		6.0		19.5		2.9		6.4		3.9		13.6		20.5		2.6	
1.1	U	1.1	U	1.1	U	1.1	U	2.4		1.1	U	1.2		1.1	U	1.9		2.4		1.1	U
0.85		1.1		0.61		1.7		8.8		0.92		2.4		1.5		4.4		15.4		0.84	
2.4		2.1		1.9		3.9		15.0		2.3		6.2		3.3		9.3		15.7		2.3	
3.4		3.8		1.9		6.5		45.1		3.3		8.0		5.8		25.5		96.0		3.7	
2.0		2.3		1.4		3.8		15.8		2.1		4.7		3.5		10.7		28.9		2.3	
3.9		4.4		2.8		7.7		21.6		3.2		10.4		5.3		19.8		30.9		3.2	
1.8		2.4		1.6		4.1		14.9		1.8		5.3		3.2		7.6		16.4		1.9	
2.3		2.4		2.0		4.0		14.0		2.3		5.5		4.2		10.7		18.0		2.2	
6.5		6.8		4.8		12.3		38.2		6.5		13.5		10.3		25.6		34.3		6.1	
0.50	U	1.5		0.50	U	0.98		9.0		0.50	U	1.0		0.92		4.8		16.0		0.65	
2.2		2.7		1.7		4.9		19.8		2.2		6.5		3.7		11.5		35.9		2.5	
2.0		1.6		1.5		4.1		13.2		1.9		5.6		3.1		7.9		18.9		1.9	
1.9		2.2		1.4		4.1		23.3		2.2		4.9		3.6		14.2		36.1		2.1	
2.2		2.6		1.8		5.1		11.4		2.6		4.9		3.7		10.5		18.8		2.4	
0.77		0.86		0.58	U	1.6		5.3		0.7		2.3		1.4		4.5		7.3		0.62	
1.4		4.0		1.3		3.8		19.9		1.2		3.6		2.7		11.1		29.6		1.8	
0.78	U	0.78	U	0.78	U	0.78	U	4.2		0.78	U	0.78	U	0.78	U	2.3		8.9		0.78	U
0.66	U	1.6		0.66	U	1.42		13.0		0.66	U	1.4		1.0		4.7		17.7		0.84	
1.9		2.7		1.6		4.7		20.0		1.7		4.3		3.7		14.4		38.1		2.2	
1.6		1.9		1.4		3.2		14.2		1.9		4.4		3.0		10.7		20.0		1.5	
1.2		1.7		1.0		2.4		14.1		1.3		3.9		3.0		8.4		17.1		1.4	
0.88		0.97		0.65		1.7		5.9		0.95		2.2		1.5		4.6		7.6		0.77	
1.08		1.2		0.72	U	2.0		6.0		0.74		2.4		1.4		3.8		5.6		0.81	
2.20		2.6		1.9		4.8		22.4		2.0		6.3		4.3		14.0		34.8		2.2	
0.46		0.64		0.26	U	1.1		9.3		0.43		1.0		0.79		4.2		17.1		0.61	
0.36	U	0.36	U	0.36	U	0.54		3.1		0.36	U	0.49		0.36	U	1.6		5.8		0.36	U
1.1		1.5		0.73		2.0		6.7		0.99		2.6		1.6		5.1		10.9		0.89	
1.0		1.6		0.87		2.6		11.2		1.1		3.2		2.2		6.1		20.1		1.2	
2.8		3.2		2.59		6.59		20.83		3.47		7.37		4.4		15.7		23.5		3.4	
2.0		2.7		1.77		4.55		20.20		2.04		5.57		4.1		10.9		31.3		2.5	
2.6		3.1		2.01		5.54		20.30		2.20		6.03		4.9		14.8		22.9		2.3	
11.8		11.3		8.45		18.67		65.55		9.52		27.48		17.8		47.1		88.7		10.4	
3.4		4.2		2.92		7.14		26.77		3.10		11.34		6.0		18.5		36.5		4.1	
4.1		4.1		3.25		5.79		25.26		4.25		9.21		6.4		19.2		27.6		3.4	

Table 12.3

Continued

Sample	PCB 56	Qualifier	PCB 60	Qualifier	PCB 64	Qualifier	PCB 66	Qualifier	PCB 70	Qualifier	PCB 71	Qualifier	PCB 74	Qualifier	PCB 77	Qualifier
Sample 1	4.1		1.9		1.9		8.7		20.0		0.66		5.2		0.52	
Sample 2	2.8		2.0		2.0		6.3		10.5		0.82		3.9		0.66	U
Sample 3	3.1		2.5		2.8		6.5		7.7		1.3		4.9		0.46	
Sample 4	6.4		5.9		7.3		17.7		21.6		3.2		8.9		1.1	
Sample 5	11.1		5.5		4.5		16.0		53.0		2.0		13.7		1.1	
Sample 6	6.3		5.1		7.0		13.4		20.0		3.8		9.6		1.2	
Sample 7	15.4		14.8		15.9		39.5		42.5		9.5		22.1		2.7	
Sample 8	11.9		8.3		11.6		21.5		35.6		7.0		15.1		1.7	
Sample 9	1.5		1.3	U	1.3	U	2.8		4.8		1.3	U	1.7		1.3	U
Sample 10	9.7		6.2		8.7		20.4		27.7		4.9		15.0		1.5	
Sample 11	12.5		9.8		7.3		32.3		49.7		2.8		18.7		2.1	
Sample 12	7.1		5.9		7.3		21.2		24.5		5.1		12.9		1.5	
Sample 13	13.1		10.3		10.0		25.7		36.7		6.1		17.5		1.6	
Sample 14	13.4		7.8		6.4		27.4		46.8		2.6		18.6		1.2	
Sample 15	8.7		7.6		8.7		16.8		25.1		3.9		11.1		1.2	
Sample 16	12.9		9.9		10.6		27.6		42.8		5.1		15.5		1.7	
Sample 17	2.2		1.9		1.6		4.6		5.4		1.1	U	2.8		1.1	U
Sample 18	8.6		5.1		4.0		14.3		35.1		1.5		12.5		0.83	
Sample 19	12.1		9.4		8.8		24.8		36.7		4.9		19.1		1.4	
Sample 20	16.3		8.3		18.5		28.1		66.5		6.9		28.1		1.7	
Sample 21	8.0		6.4		7.8		20.3		24.6		4.5		15.7		1.1	
Sample 22	18.6		13.3		13.3		36.2		60.9		7.4		27.6		1.6	
Sample 23	9.2		6.3		7.2		17.1		25.1		3.9		11.4		1.1	
Sample 24	10.1		7.3		10.5		28.0		38.0		4.7		18.2		1.8	
Sample 25	25.1		20.6		23.2		49.2		67.8		13.6		33.4		4.2	
Sample 26	10.2		5.9		3.6		22.0		49.7		1.2		15.3		1.1	
Sample 27	13.7		8.0		8.6		31.8		48.8		5.0		20.0		2.0	
Sample 28	7.0		5.7		7.5		18.4		19.7		4.6		11.4		1.1	
Sample 29	9.2		5.9		8.8		22.1		37.8		3.7		16.0		1.1	
Sample 30	9.8		7.2		10.0		17.2		22.0		4.3		13.4		1.5	
Sample 31	4.3		2.8		3.1		8.4		11.0		1.8		5.1		0.58	U
Sample 32	26.1		20.0		10.3		62.3		98.3		3.8		33.6		3.4	
Sample 33	3.2		1.9		1.9		6.0		13.5		0.78	U	3.6		0.78	U
Sample 34	13.5		9.0		5.5		35.2		70.0		1.8		19.6		1.7	
Sample 35	15.4		9.4		11.0		33.0		54.2		4.1		20.0		1.8	
Sample 36	9.1		6.6		8.1		18.9		22.9		3.2		11.4		1.2	
Sample 37	6.7		5.7		5.1		17.0		22.2		3.0		9.4		1.0	
Sample 38	4.1		2.9		3.5		9.6		11.4		2.1		5.5		0.62	U
Sample 39	5.2		3.2		3.3		9.3		15.2		2.1		5.9		0.75	
Sample 40	8.7		7.1		10.8		19.2		29.2		4.7		13.3		1.2	
Sample 41	4.4		2.2		2.7		9.2		19.7		0.87		6.0		0.55	
Sample 42	1.1		0.64		1.1		1.8		4.6		0.36		1.7		0.36	U
Sample 43	6.2		4.9		4.8		11.5		22.4		2.0		8.5		0.8	
Sample 44	11.0		7.1		5.8		19.9		30.0		2.5		10.6		1.2	
Sample 45	14.8		10.0		13.0		28.1		46.0		6.8		20.5		2.9	
Sample 46	16.3		8.9		10.7		34.0		55.3		5.1		21.2		2.3	
Sample 47	11.6		8.9		10.1		22.5		31.1		5.5		17.9		1.6	
Sample 48	32.1		28.3		33.6		93.1		96.4		22.7		67.4		5.5	
Sample 49	20.5		15.1		14.4		48.2		59.9		8.0		31.6		2.9	
Sample 50	13.4		13.1		16.5		32.4		40.0		8.3		23.7		2.3	

PCB 82	Qualifier	PCB 84	Qualifier	PCB 85	Qualifier	PCB 87	Qualifier	PCB 92	Qualifier	PCB 95	Qualifier	PCB 97	Qualifier	PCB 99	Qualifier	PCB 101	Qualifier	PCB 105	Qualifier	PCB 110	Qualifier
5.0		6.7		6.1		14.0		3.0		12.2		9.9		14.3		27.3		17.0		33.2	
1.7		3.3		3.1		6.0		1.5		8.9		5.6		6.1		13.3		8.5		16.4	
1.2		2.0		1.8		3.0		0.80		3.6		2.7		3.7		5.8		3.6		6.9	
2.4		3.1		3.7		4.7		1.2		5.5		4.2		6.3		9.0		6.7		11.0	
10.8		19.8		15.7		40.2		9.1		47.2		23.7		37.6		65.7		51.0		88.3	
3.4		3.7		4.7		5.4		1.6		7.1		4.6		8.9		12.0		8.5		15.2	
9.0		16.3		12.4		24.4		7.2		38.8		17.1		24.4		45.9		27.5		48.6	
7.7		11.1		9.4		20.9		5.4		27.6		13.3		18.5		39.2		19.6		44.4	
1.3		1.9		1.6		3.0		1.3	U	3.4		2.2		3.4		6.3		4.0		6.7	
3.6		4.6		6.7		8.5		1.7		6.1		5.8		11.7		16.1		15.3		19.2	
11.6		10.9		12.3		20.7		3.7		16.6		14.1		25.8		39.7		42.6		47.6	
3.5		6.5		5.5		9.2		3.1		13.5		7.3		10.4		20.0		12.5		27.8	
5.3		12.0		8.2		17.2		4.3		23.0		10.3		17.3		31.8		16.7		37.5	
9.8		12.5		14.3		21.9		4.3		14.7		16.7		27.6		36.9		39.0		56.4	
4.6		8.3		6.1		11.5		3.0		14.5		9.8		10.3		17.3		13.8		27.1	
6.6		8.7		12.2		16.3		3.3		12.6		12.1		17.8		29.1		23.7		35.4	
1.1	U	1.1	U	1.6		2.3		1.1	U	1.6		1.7		3.0		3.5		3.1		5.2	
6.04		8.22		10.74		14.49		3.84		14.25		11.2		17.9		29.7		30.5		43.3	
6.6		6.9		9.9		11.4		2.4		11.7		10.5		21.2		19.5		22.5		31.7	
19.3		42.8		23.4		65.3		20.0		106.9		45.5		57.5		115.2		51.1		128.4	
5.2		8.1		5.9		13.1		3.9		19.1		10.6		11.7		25.9		15.2		32.5	
5.5		12.1		14.3		15.8		4.3		23.2		17.7		24.7		33.1		41.5		27.9	
3.8		5.9		5.5		10.9		2.5		9.9		6.3		10.7		18.7		10.8		18.3	
6.2		8.0		9.6		11.7		2.1		11.2		10.4		13.8		19.3		25.4		28.3	
8.1		11.5		16.7		22.0		4.5		19.5		19.1		27.3		35.0		35.9		51.7	
11.3		14.3		12.5		23.5		6.1		20.2		22.1		31.7		45.0		40.1		67.4	
10.6		16.3		13.7		24.7		6.3		31.8		19.2		30.8		50.1		36.3		74.2	
3.2		5.2		4.1		9.2		2.4		10.3		5.8		8.5		13.0		8.1		19.8	
9.3		16.2		11.6		25.1		7.8		30.6		19.6		23.6		51.0		30.0		51.3	
2.9		5.0		3.3		6.6		2.2		10.0		5.0		8.8		13.1		8.9		18.8	
1.3		2.8		2.7		4.7		1.1		5.7		3.4		5.4		7.7		6.0		8.8	
21.7		26.1		41.5		62.2		12.1		39.5		38.7		68.2		91.4		112.6		137.7	
3.5		3.9		4.8		8.3		1.8		10.0		5.5		9.8		15.8		11.0		19.4	
13.9		19.3		23.3		37.3		7.3		23.3		28.3		44.3		53.2		68.4		83.2	
13.5		16.1		14.4		34.6		8.8		34.3		20.2		38.8		68.4		42.8		72.8	
4.0		7.9		7.0		13.2		3.6		12.5		8.1		12.7		21.6		16.2		33.2	
4.9		6.1		5.8		13.0		3.1		15.1		8.2		12.7		21.2		14.4		26.7	
1.9		2.8		2.9		4.8		1.0		5.6		3.4		5.1		7.6		6.4		11.8	
2.3		2.5		4.1		5.3		1.0		3.9		4.8		8.0		9.5		8.7		13.0	
6.3		11.9		8.9		16.5		4.8		24.2		12.6		12.9		31.6		14.2		36.4	
4.2		7.3		6.2		14.1		3.8		19.9		11.8		12.0		28.6		16.4		30.4	
1.4		2.3		1.5		3.4		1.2		6.1		2.3		3.7		7.6		3.2		8.5	
3.9		6.2		6.1		9.9		2.0		9.6		8.7		9.8		19.4		12.3		27.0	
6.9		7.4		11.9		15.7		4.5		18.7		10.4		19.9		26.2		30.6		27.2	
6.0		8.8		10.1		12.2		3.0		10.6		11.8		19.3		18.8		20.6		33.3	
11.3		12.9		16.8		26.5		5.9		32.1		20.3		33.8		50.4		40.8		65.7	
5.6		8.9		9.8		15.5		4.2		16.0		10.2		15.2		29.8		16.4		40.1	
12.4		17.1		15.9		27.2		7.2		37.2		26.0		33.4		50.4		30.3		57.3	
9.8		17.0		18.4		28.5		5.7		28.0		18.5		38.5		59.5		48.6		70.7	
5.6		6.2		8.0		10.8		2.4		12.4		10.6		14.8		20.3		19.4		26.3	

Table 12.3
Continued

Sample	PCB 118	Qualifier	PCB 128	Qualifier	PCB 130	Qualifier	PCB 132	Qualifier	PCB 135	Qualifier	PCB 136	Qualifier	PCB 137	Qualifier	PCB 138	Qualifier
Sample 1	34.3		4.9		2.1		6.9		1.6		1.5		1.7		24.9	
Sample 2	16.7		2.2		0.93		3.3		0.72		0.81		0.66		6.9	
Sample 3	6.1		0.87		0.44	U	1.3		0.44	U	0.44	U	0.44	U	3.4	
Sample 4	9.6		1.5		0.92	U	1.6		0.92	U	0.92	U	0.92	U	6.0	
Sample 5	90.2		14.5		5.3		17.6		4.6		4.8		3.7		42.5	
Sample 6	18.8		2.1		0.98	U	2.1		0.98	U	0.98	U	0.98	U	8.2	
Sample 7	47.0		7.1		3.4		11.1		3.0		3.9		2.3		34.4	
Sample 8	37.6		5.5		2.3		8.5		2.7		3.3		1.8		26.7	
Sample 9	8.1		1.3	U	1.3	U	1.6		1.3	U	1.3	U	1.3	U	5.0	
Sample 10	27.4		3.0		0.81		2.7		0.63		0.61		0.87		10.0	
Sample 11	78.1		11.2		3.1		11.0		2.0		1.6		3.0		39.1	
Sample 12	26.4		3.1		1.0		4.7		1.2		1.4		0.88		11.8	
Sample 13	29.3		5.2		2.0		6.2		1.9		2.2		1.6		16.8	
Sample 14	78.1		9.7		3.3		8.7		2.1		1.5		2.7		34.4	
Sample 15	20.9		4.4		1.2		6.4		1.3		1.8		1.2		15.4	
Sample 16	44.9		5.6		2.0		5.2		1.5		1.6		1.6		21.7	
Sample 17	5.6		1.1	U	1.1	U	1.1	U	1.1	U	1.1	U	1.1	U	2.6	
Sample 18	55.3		7.2		2.5		6.5		2.0		1.7		2.2		26.5	
Sample 19	40.6		5.6		1.3		5.5		0.96		1.1		1.4		16.2	
Sample 20	126.1		25.1		8.8		39.0		8.1		7.0		7.1		103.7	
Sample 21	27.5		4.4		1.8		8.1		2.0		1.7		1.5		18.7	
Sample 22	9.4		7.7		25.3		7.4		0.29		7		1		2	
Sample 23	26.6		2.8		1.2		4.4		1.1		1.0		1.0		13.1	
Sample 24	41.3		4.7		1.4		4.5		1.0		0.94		1.5		15.6	
Sample 25	60.0		6.0		2.1		8.6		1.7		1.4		2.1		26.3	
Sample 26	94.0		12.0		3.4		11.9		3.1		2.8		3.7		43.4	
Sample 27	77.8		9.8		4.2		16.2		3.6		3.6		2.7		41.4	
Sample 28	16.3		2.1		0.78		3.3		0.8		1.1		0.8		9.4	
Sample 29	53.7		7.3		3.7		15.0		4.7		3.5		2.4		35.1	
Sample 30	17.8		2.0		0.68		2.8		0.75		0.94		0.47		7.8	
Sample 31	11.3		1.5		0.58	U	1.6		0.58	U	0.58	U	0.58	U	5.2	
Sample 32	213.0		23.8		8.6		26.4		4.9		4.9		6.2		87.6	
Sample 33	19.3		3.3		1.4		3.7		1.1		1.1		0.96		11.7	
Sample 34	110.8		13.2		5.4		14.8		3.5		3.5		4.5		59.1	
Sample 35	109.7		14.5		4.5		15.6		4.6		4.1		3.1		57.1	
Sample 36	30.6		5.5		1.5		6.9		1.3		1.6		1.1		19.3	
Sample 37	33.4		5.0		1.5		5.4		1.2		1.5		1.4		14.9	
Sample 38	11.4		1.5		0.62	U	1.86		0.62	U	0.62	U	0.62	U	5.1	
Sample 39	17.8		2.5		0.72	U	2.34		0.72	U	0.72	U	0.72	U	7.4	
Sample 40	34.6		6.5		2.2		9.0		2.5		2.5		1.9		22.2	
Sample 41	36.9		5.7		2.3		8.5		1.9		2.2		1.8		20.1	
Sample 42	8.2		1.3		0.55		1.9		0.58		0.63		0.39		5.55	
Sample 43	32.8		3.8		1.2		4.6		0.92		1.0		1.3		13.2	
Sample 44	52.7		7.5		2.4		9.0		1.7		2.0		2.5		24.2	
Sample 45	38.2		5.3		1.6		5.4		0.84		0.9		1.5		18.3	
Sample 46	74.9		11.0		3.9		13.5		2.9		2.8		3.6		39.4	
Sample 47	40.2		4.9		1.6		6.1		1.6		1.6		1.3		20.5	
Sample 48	67.6		7.1		2.7		10.4		3.1		3.0		2.1		24.7	
Sample 49	94.5		10.8		3.1		12.4		2.9		3.2		3.3		37.2	
Sample 50	31.4		3.6		0.98		4.4		0.88		1.0		1.0		12.8	

PCB 141	Qualifier	PCB 146	Qualifier	PCB 149	Qualifier	PCB 151	Qualifier	PCB 153	Qualifier	PCB 156	Qualifier	PCB 158	Qualifier	PCB 163	Qualifier	PCB 170	Qualifier	PCB 180	Qualifier
10.5		2.0		10.3		1.4		12.1		3.4		3.2		2.7		1.8		2.1	
3.6		0.9		3.8		0.80		5.2		1.2		1.1		1.5		0.76		1.19	
2.2		0.44	U	1.7		0.44	U	1.8		0.57		0.53		0.59		0.44	U	0.46	
4.2		0.92	U	2.5		0.92	U	3.1		0.92	U	0.92	U	0.92	U	0.92	U	0.92	U
26.5		5.4		30.3		4.2		25.4		9.6		6.9		8.5		4.3		5.9	
6.0		0.98	U	4.0		0.98	U	3.94		1.16		1.18		1.30		0.98	U	1.0	
22.4		3.8		19.0		3.4		19.1		5.0		4.6		4.8		2.8		3.7	
15.7		3.2		14.9		3.0		18.2		3.8		3.8		4.6		2.2		3.5	
5.7		1.3	U	2.3		1.3	U	2.4		1.3	U	1.3	U	1.3	U	1.3	U	1.3	U
5.0		0.97		4.8		0.69		7.4		2.1		1.4		1.6		0.76		1.6	
16.9		2.7		13.8		1.6		20.4		6.3		4.6		5.0		2.3		3.3	
10.5		1.0		6.6		1.3		7.7		1.9		1.7		2.1		1.2		1.4	
10.0		2.5		9.7		2.3		15.2		2.9		2.5		3.4		1.7		2.9	
26.1		2.3		12.5		2.1		20.3		7.2		5.2		3.5		2.2		3.0	
9.2		1.6		7.9		1.6		10.6		1.9		2.0		2.5		1.4		1.9	
13.2		1.9		9.3		1.5		14.1		3.4		2.8		3.6		1.4		2.9	
6.0		1.1	U	1.2		1.1	U	1.6		1.1	U	1.1	U	1.1	U	1.1	U	1.1	U
13.8		2.5		11.1		1.6		15.0		4.7		3.8		3.3		1.7		2.6	
8.6		1.6		5.7		0.9		9.8		3.2		2.4		2.7		1.2		1.5	
57.6		10.0		62.8		9.7		59.6		13.9		10.7		11.6		8.5		11.1	
12.1		1.9		11.5		2.2		11.0		2.4		2.8		3.4		1.8		2.4	
7		17		6		0.97		17.76		4.59		5.22		3.59		2.45		2.87	
9.9		1.3		6.6		1.2		8.2		2.0		2.2		2.0		1.0		1.2	
12.2		1.3		7.4		1.1		9.4		3.5		3.0		2.0		1.2		1.5	
12.2		2.4		9.3		1.9		15.4		3.7		3.5		3.2		1.8		3.1	
31.1		3.5		20.0		2.7		25.0		7.6		6.2		6.5		2.6		4.0	
32.6		3.6		17.5		4.2		25.9		6.6		5.3		6.9		3.5		3.8	
4.4		1.1		5.6		0.98		7.0		1.8		1.3		1.9		0.79		1.3	
17.0		3.6		19.9		4.0		23.3		6.4		5.6		6.4		3.5		3.6	
6.8		0.91		4.5		0.93		4.1		1.1		1.1		1.4		0.86		1.3	
2.5		0.58	U	2.7		0.58	U	3.5		0.85		0.82		0.94		0.58	U	0.60	
58.9		8.6		34.4		5.1		51.4		17.0		13.6		12.3		5.5		7.5	
9.1		1.3		5.5		0.93		8.5		2.3		2.0		1.9		1.0		1.4	
35.4		5.1		22.3		2.3		37.7		11.3		8.9		7.3		3.8		3.7	
43.3		4.7		26.3		4.6		30.5		7.9		6.9		8.0		3.5		4.8	
2.8		1.6		8.7		1.5		9.7		3.0		2.7		2.4		1.6		1.9	
2.3		1.8		8.1		1.5		10.7		2.6		2.5		2.4		1.2		1.8	
0.93		0.62	U	2.7		0.62	U	4.9		0.86		0.66		0.94		0.62	U	0.74	
0.98		0.72	U	3.3		0.72	U	4.6		1.2		1.1		0.91		0.72	U	0.80	
4.3		2.4		16.1		2.5		15.8		3.4		3.0		4.2		2.5		3.2	
3.3		2.2		10.1		2.1		14.0		3.0		3.3		3.1		1.9		2.5	
0.95		0.54		3.0		0.76		3.5		0.75		0.75		1.0		0.40		0.66	
1.7		1.2		6.1		0.99		9.0		2.7		2.1		2.0		1.1		1.4	
3.3		2.2		9.1		2.2		14.5		4.8		3.8		3.4		1.9		2.6	
2.1		1.3		5.7		0.95		9.6		3.4		2.5		2.7		1.4		1.8	
5.0		3.6		17.8		3.1		25.0		7.4		6.0		6.0		3.6		3.5	
3.0		2.5		11.5		1.9		11.1		3.2		2.8		2.8		1.7		2.2	
5.2		3.5		20.0		3.0		22.9		4.5		3.5		5.5		2.9		5.6	
5.8		3.4		21.4		2.6		21.5		6.4		5.0		5.6		2.9		3.8	
1.8		1.3		6.2		1.0		9.8		2.4		2.1		1.7		1.2		1.9	

Two principal components account for more than 92% of the variance in Data Set 1, and the scores plot clearly divides the samples into two clusters: Aroclor 1248 and Aroclor 1254.

There is an obvious problem though. While this is indeed a two Aroclor system (as Figure 12.2 clearly suggests) it is not a two-source system. The first two PCs do not differentiate between the two Aroclor 1254 variants. This illustrates a common problem in the application of PCA to environmental chemical data. All too often, investigators will present a two-PC scores plot like Figure 12.2, accompanied by a statement of justification indicating that two principal components accounts for 92.5% of the variance. Such a statement leaves the tacit implication that the residual 7.5% of the variance is random noise, which in this case is clearly not the case.

A three component scores plot for Data Set 1 is shown as Figure 12.3. Three PCs account for 97.5% of the variance; an incremental increase over the percentage accounted for by two PCs. However, that small percentage of total variance is *not* random. The three-PC scores plot clearly distinguishes three clusters, rather than two, and effectively allows the analyst to infer the presence of three sources.

This example highlights several important precautionary notes.

1 For better or for worse, the use of mathematical techniques such as PCA carries with it the aura of precision and exactitude. In the case of the two-PC scores plot (Figure 12.2) the strong clustering into two Aroclor groups, coupled with the statement that two PCs account for 92.5% of the variance, may be sufficiently intimidating to impress skeptics. Moreover, it may even provide a false sense of security to the naïve analyst.

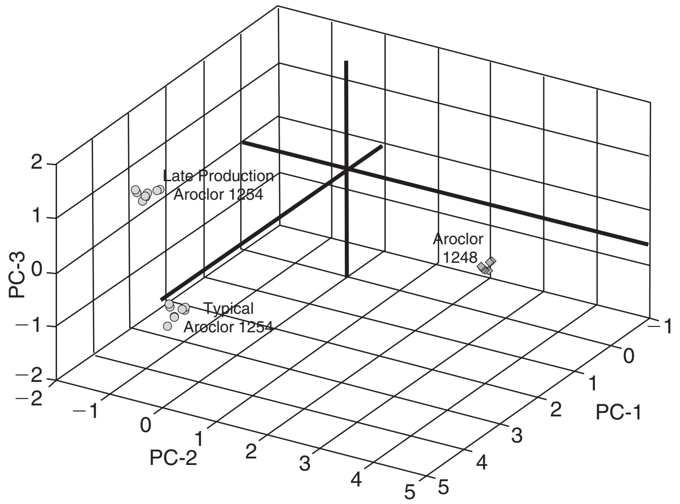


Figure 12.3
Three principal component scores plot of Data Set 1. Three PCs account for 97.5% of the variance, and allow clear distinction of the three PCB sources.

2 In an environmental forensics setting, we seldom have prior knowledge of the true data structure, so we should not be arrogant in our application of rules-of-thumb regarding what percentage of variance should be considered 'significant'.

3 Scientific and legal 'significance' is clearly *not* a function of variance. If a party involved in environmental litigation had used Aroclor 1254, but ceased all operations in the mid-1960s (prior to Monsanto's change in the Aroclor 1254 production process), the small difference in percentage of variance between two and three PCs would have enormous implications.

4 Practitioners of PCA-based methods in environmental forensics must employ more sophisticated goodness-of-fit diagnostics than percentage variance, and they must have a reasonable understanding of how such methods work.

Because PCA is a powerful exploratory data analysis tool on its own, as well as an intermediate step in receptor modeling methods, we will discuss this method in considerable detail below. The key steps in PCA include: (1) data transformations; (2) singular value decomposition (eigenvector decomposition); (3) determination of the number of significant eigenvectors; and (4) visual display of scores and loadings plots. Each of these steps are discussed in turn below.

12.2.2 DATA TRANSFORMATIONS

To the data analyst, the laboratory results received from the chemist are 'raw data'. In environmental chemistry these raw data are usually transmitted in units of concentration. Data Set 2 (Table 12.3: presented in units of ng/g) is a good example. However, seldom is a statistical analysis performed on a matrix in this form. Rather, the matrix is transformed in some manner. A data transformation is the application of some mathematical function to each measurement in a matrix. Taking the square root of every value in a matrix is an example. In analysis of environmental chemical data, data transformations are done either for (1) reasons related to the environmental chemistry of the system; or (2) mathematical reasons, to optimize the analysis.

A transformation commonly done for chemical or physical reasons is sample normalization. In an environmental system, concentrations can vary widely due to dilution away from a source. For example, in the case of contaminated sediment investigations, concentrations may decrease exponentially away from an effluent pipe. However, if the relative proportions of individual analytes remain relatively constant then we might infer a single source scenario, coupled with dilution away from the source. Inference of source by pattern recognition techniques is concerned more with the relative proportions between analytes than with absolute concentrations. Thus, a transformation is necessary to normalize out concentration/dilution effects. One that is commonly employed is a transformation to a percent metric, where each value is divided by the total concentration of the sample. This percent transformation is also referred to as a 'constant row-sum' transformation, because the sum of analyte concentrations in each sample (i.e., across rows) sums to 100%. Stated mathematically, where **S** is an $m \times m$ diagonal matrix with the m row-sums (total concentrations) of $\mathbf{X}_{\mathbf{(nk)}}$ along the diagonal, and zeros on the off-diagonals, the constant row-sum matrix **X** may be calculated as:

$$\mathbf{X} = 100 * \mathbf{S}^{-1} * \mathbf{X}_{\mathbf{(nk)}} \tag{12.2}$$

An alternative to the constant row-sum transformation is to normalize the data with respect to a single species or compound (a so-called 'normalization variable'). This transformation involves setting the value of the normalization variable to 1.0, and the values of all other variables to some proportion of 1.0, such that their ratios with respect to the normalization variable remain the same as in the original metric.

The second type of transformation is done more for mathematical/statistical purposes. In any chemical data set, there is usually a strong relationship between the mean value of an analyte and its variance (variance is square of the standard deviation). Therefore, chemicals measured in trace concentrations almost always exhibit smaller variance than those measured at much higher concentrations. Multivariate procedures such as PCA are variance driven, so in the absence of some transformation across variables, the analytes with highest mean and variance usually have the greatest influence on the analysis. Polychlorinated dibenzo-*p*-dioxins provide a particularly instructive example. In most environmental systems, 2,3,7,8-TCDD (dioxin) is typically measured at orders of magnitude lower concentrations than the octa-chlorinated congener OCDD. Thus the mean and variance of 2,3,7,8-TCDD is typically orders of magnitude lower than OCDD. However, this does not imply less precision or accuracy in the chemical measurement of 2,3,7,8-TCDD. Nor does it imply that 2,3,7,8-TCDD is of secondary environmental importance to OCDD. In fact, the opposite is usually the case because 2,3,7,8-TCDD has much higher toxicity than does OCDD. As such, analysis of environmental chemical data almost always requires some sort of 'homogeneity of variance' transformation. A number of transformations may be applied to produce homogeneity of variance. One of the most commonly used transformations in PCA is autoscale transform (also known as the Z transform). Given a matrix **X**, with calculated means $(\bar{x}_j)$ and standard deviations (s_j) in each column $(j = 1, 2 \ldots n)$. The autoscaled matrix **Z** is calculated:

$$\mathbf{Z}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{12.3}$$

The autoscale transformation guarantees absolutely equal variance in that it sets the mean of each column to 0.0 and the standard deviation to 1.0.

Another common homogeneity of variance transformation is the range-transform (also known as the minimum/maximum transformation). Following the convention of Miesch (1976a), where the original matrix is denoted **X**, the range is denoted as X-prime ($\mathbf{X}' = \{x'_{ij}\}$). The transformation is performed as follows:

$$x'_{ij} = (x_{ij} - x_{\min j})/(x_{\max j} - x_{\min j}) \tag{12.4}$$

This results in a matrix where the minimum value in each column equals 0.0 and the maximum equals 1.0. The range transformation produces variances that are approximately homogeneous (unlike the Z transform, which results in absolute homogeneity of variance). However, the range transform has two advantages: (1) it does not force the data to conform to a theoretical distribution that might not be applicable (in the case of the Z transform: the standard normal distribution); and (2) because transformed values are within the 0.0 to 1.0 range, all values are non-negative. The latter feature is particularly useful in receptor modeling, where explicit constraints of non-negativity are used to derive source fingerprints. Malinowski (1977, 1991) notes that a disadvantage of the range transform is that it is extremely sensitive to outliers. As we will see, however, in an environmental forensics investigation, detection and evaluation of outliers is a crucial part of the process, so this feature of the range transform is usually desirable.

A third transformation (the equal vector length transform) is often, but not always, employed in multivariate analysis of chemical data. The equal vector length transform is performed to force each of the sample vectors to have equal Euclidean length. If all vectors have equal length, then the differences between samples are a function only of the angles between samples. Thus the similarities and differences between samples can be expressed as a similarity matrix of cosines (Davis, 1986). The cosine between two identical samples is 1.0 – the cosine between two completely dissimilar samples (i.e. vectors at 90°) is 0.0. By Miesch's convention (Miesch, 1976a), the constant-sum input matrix is indicated as **X**, the range transform matrix as $\mathbf{X}'$, and the equal vector length transform as $\mathbf{X}''$. For matrix algebra-based programming languages, a computationally efficient way (Hopke, 1989) is to first define $\mathbf{Y} = \{y_{ij}\}$ (the matrix element of **Y** in row i, column j). **Y** is a $m \times m$ diagonal matrix where each diagonal element equals the inverse of the square root of the sum-of-squares along rows of $\mathbf{X}'$:

$$y_{ij} = \frac{1}{\left(\sum_{i=1}^{m} x'^{\,2}_{ij}\right)^{1/2}} \tag{12.5}$$

The transformed matrix $\mathbf{X}''$ may then be calculated as follows:

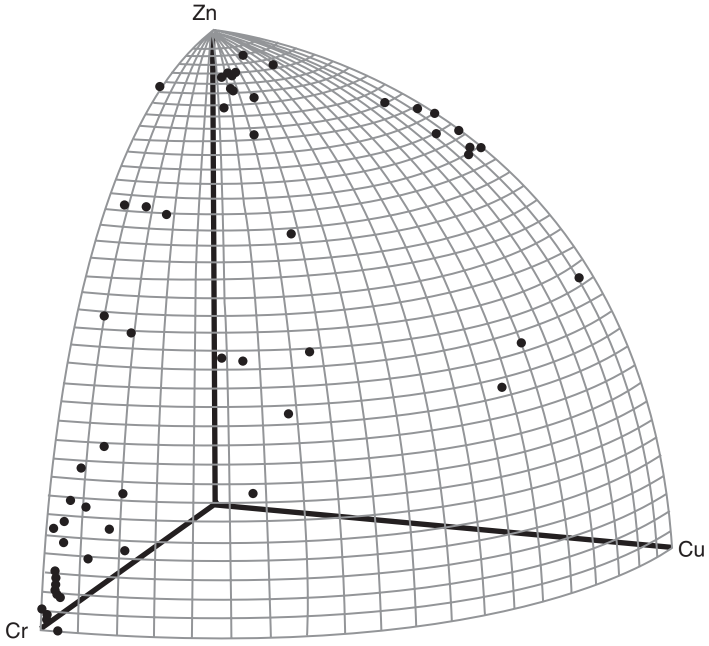
$$\mathbf{X}'' = \mathbf{Y}\mathbf{X}' \qquad (12.6)$$

This transformation has an added advantage for receptor modeling. By definition, if all samples have equal vector lengths, each sample must lie on an $n - 1$ dimensional surface, unit length from the origin. To demonstrate this, a simple three chemical example is shown in Figure 12.4. In this case $n = 3$, because only three variables (chromium, copper and zinc) are present. These data were transformed by Equations 12.5 and 12.6; thus all samples lie on a two-dimensional surface, which is unit length from the origin ($n - 1 = 2$). As we will see in Section 12.3, receptor/mixture modeling involves resolution of a $k - 1$ dimensional geometric figure within k dimensional principal component space. Thus, this transformation has particularly attractive features in that regard.

In summary, normalization across rows (constant row sum or normalization to a marker chemical) is almost always done in environmental chemometric analyses. Some homogeneity of variance transformation (e.g., range transform or autoscale transform) is also typically performed in order to keep high concentration variables from dominating an analysis. The constant vector length transform is sometimes done, usually when (1) there is some advantage to being able to express relationships between samples simply in terms of angles between samples; or (2) when the PCA is an intermediate step in receptor

Figure 12.4

A simple three chemical system (n = 3: chromium, copper and zinc) illustrates the two-dimensional (n − 1) geometry of a matrix which has undergone the equal vector length transformation. All sample vectors are unit length from the origin, and lie on the n − 1 dimensional spherical surface.



modeling, which involves resolution of a $k - 1$ dimensional simplex within k dimensional space. The PCA that resulted in the plots shown on Figures 12.2 and 12.3 involved transformation by the constant row sum transformation (Equation 12.2) followed by the range transform (Equation 12.3). The constant vector length transformation was not used.

12.2.3 EIGENVECTOR DECOMPOSITION

Eigenvector decomposition is a simple mathematical procedure that allows a reduction in dimensionality of a data set. This is the core mathematical operation involved in principal components analysis. It is most often accomplished through singular value decomposition (SVD) of the transformed matrix $\mathbf{X}'$ or $\mathbf{X}''$. As shown in Figure 12.4, the transformed $m \times n$ matrix may be thought of as m vectors plotted in n dimensional space. Each variable is represented as one of n orthogonal axes of a cartesian coordinate system. There are, however, an infinite number of sets of n mutually orthogonal basis vectors that may equivalently be used to plot the sample vectors, without loss of information. The eigenvectors extracted from a similarity matrix of the original data is one such alternative reference space. The number of eigenvectors (i.e., the number of principal components) will equal m or n, whichever is smaller. However, there are usually correlations between analytes due to common sources. Thus, a relatively small subset of the eigenvectors is typically sufficient to capture the variability in the system, and the interrelationships between samples can be observed without loss of information.

Eigenvector decomposition is a well-established part of the core knowledge of mathematics and is frequently used in the physical and natural sciences. The calculation of eigenvectors and eigenvalues is relatively straightforward, but lengthy and cumbersome. As such, a conceptual discussion of the topic is presented below, and the reader is referred to any number of elementary linear algebra texts for a complete mathematical discussion. Davis (1986) provides a detailed, lucid but less rigorous treatment, using examples from the earth sciences.

Given an error free, noise free matrix of $m > k$ samples and $n > k$ variables, resulting from k sources, only k nonzero eigenvectors and eigenvalues will be extracted. If $k = 3$, the first eigenvector will account for a high percentage of the total variance in the data set. The second eigenvector is constrained in that it must be mutually orthogonal to the first, and accounts for the highest percentage of residual variance (that variance not accounted for by the first eigenvector). The third eigenvector is mutually orthogonal to the first two, and accounts for the remainder of the variance. The data set may equivalently be expressed in this three-dimensional reference space, without loss of information. Given a transformed matrix $\mathbf{X}'$ composed of m samples along the rows, and n variables (chemical analytes) along columns, PCA is accomplished through SVD of $\mathbf{X}'$:

$$\mathbf{X}' = \underbrace{\overbrace{\mathbf{U}}^{\mathbf{F}'_R}\;\mathrlap{\overbrace{\phantom{\mathbf{\Lambda}^{1/2}\;\mathbf{V}^t}}^{\mathbf{A}'_R}}\mathbf{\Lambda}^{1/2}}_{\mathbf{A}'_Q}\;\underbrace{\mathbf{V}^t}_{\mathbf{F}'_Q} \tag{12.7}$$

where $\mathbf{U}$ equals the matrix of principal components scores in R mode, $\mathbf{V}^t$ equals the matrix of scores in Q mode, and $\mathbf{\Lambda}$ equals the diagonal matrix of eigenvalues. Principal component scores ($\mathbf{F}$) and loadings ($\mathbf{A}$) matrices for R and Q modes may result by re-expression of Equation 12.7, as indicated (Zhou *et al.*, 1983). Eigenvectors are abstract in that they usually cannot be interpreted in terms of real world phenomena (although many have tried). The total number of calculated PCs equals m or n, whichever is smaller. A model involving a reduced number of principal components (k) may be represented as follows (Q mode):

Matrix dimensions

$$\underset{(m\times n)}{\mathbf{X}''} = \underset{(m\times k)}{\mathbf{A}''}\;\underset{(k\times n)}{\mathbf{F}''} + \boldsymbol{\varepsilon} \tag{12.8}$$

where k equals the number of PCs retained for the model, and ε represents error.

12.2.4 DETERMINING THE NUMBER OF SIGNIFICANT PRINCIPAL COMPONENTS

The choice of the number k is equivalent to the decision on the number of 'significant' principal components. Of the many aspects of PCA-based methods, no topic has created more argument or controversy than the criteria used to determine the correct number of eigenvectors (i.e., k, the number of factors, sources, subpopulations or end-members).

Numerous methods have been proposed for determination of k (Cattell, 1966; Exner, 1966; Malinowski, 1977; Miesch, 1976a; Wold, 1978; Ehrlich and Full, 1987; Henry *et al.*, 1999). The spirit and intent of these methods are similar: the estimated data set, as back-calculated from reduced dimensional space (i.e., $\mathbf{X}_{\text{hat}}$ or $\hat{\mathbf{X}}$), should reproduce the measured data ($\mathbf{X}$) with reasonable fidelity.

12.2.4.1 Single Index Methods

Six criteria commonly used in environmental chemometrics were applied to Data Sets 1 and 2, and the results are shown in Tables 12.4 and 12.5, respectively.

The PCA of both data sets involved the constant sum and range transformations. Each of the six indices, and the rationale for their use, is discussed below.

1 *Cumulative percentage variance.* The rationale for this criterion is simple: a reduced dimensional model should account for a large percentage of the variance in the original matrix. However, as discussed in Section 12.2.1, any *a priori* choice regarding what percentage of variance one should consider to be 'significant' is problematic. Workers in multivariate statistics, chemometrics and mathematical geology generally acknowledge that any proposed cutoff criterion is arbitrary (Malinowski, 1991; Deane, 1992; Reyment and Jöreskog, 1993). The lack of a clear criterion makes the cumulative variance method dubious. Nonetheless, in Tables 12.4 and 12.5, we note that the commonly used cutoff of 95% suggests retention of three principal components for both Data Sets 1 and 2.

2 *Scree test.* The scree test of Cattell (1966) is based on the supposition that the residual variance, not accounted for by a *k* principal component model, should level off at the point where the principal components begin accounting for random error. When residual variance is plotted versus principal component number, the point where the curve begins to level off should show a noticeable inflection point, or 'knee'. The problem with this criterion is that there is often no unambiguous inflection point, and when such is the case, the decision as to the number of significant principal components is arbitrary.

3 *Normalized varimax loadings.* The multivariate statistical algorithms of Klovan and Miesch (1976) included a subroutine that calculates the number of samples with normalized varimax loadings greater than 0.100. Neither Miesch (1976a,b) nor Klovan and Miesch (1976) included explicit discussion of its utility, but Ehrlich and Full (1987) later presented such a discussion. If an eigenvector carries systematic information, then typically, a large number of samples will have high loadings (loadings in *Q*-mode terminology). High numbered factors that account for noise and little variance typically have loadings that are small for all samples (i.e., <0.1). Factors with many samples >0.1 indicate principal components that should be retained for a model. This criterion calls for rotation of principal components using the varimax procedure of Kaiser (1958), normalization of the varimax loadings matrix to sum to 1.0 across all sample rows, and tabulation of the number of samples that exceed 0.1 for each factor. The analyst looks for a sharp drop in the index as an indication of the appropriate number of eigenvectors.

4 *Malinowski indicator function.* Malinowski (1977) presented an indicator function, which is calculated as a function of the residual standard deviation (Malinowski, 1977; Hopke, 1989). The function reaches a minimum when the 'correct' number of principal components are retained. The index has worked well with relatively simple data structures but Hopke (1989) reports that it has not proven as successful with complex environmental chemical data.

5 *Cross-validation.* Cross-validation is a commonly used method for determination of number of significant principal components. It involves successive deletion of data points, followed by prediction of the deleted data with increasing numbers of eigenvectors. The PRESS statistic (predicted residual error sum of squares) is then calculated for each number of potential eigenvectors. Many criteria have been proposed based on calculation of some function of PRESS (Wold, 1978; Eastman and Krzanowski, 1982; Deane, 1992; Grung and Kvalheim, 1994). The PRESS value and criterion in Table 12.3 is that presented by Deane (1992).

6 *Signal-to-noise ratio.* This method has been proposed very recently by Henry *et al.* (1999). Henry's NUMFACT criterion involves calculation of a signal-to-noise (S/N) ratio. Henry found that given random data, an S/N ratio as high as 2.0 could be obtained. Based on that, the rule-of-thumb criterion recommended by Henry is that principal components with S/N ratios greater than 2.0 should be retained for a model.

As is evident in Table 12.4, for Data Set 1 (a relatively simple data set with random Gaussian noise) each of these indices provides an accurate estimate of the true number of Aroclor sources. All six indices correctly indicate a three-component system. Table 12.5 reports the values for the same indices, as applied to the more complicated, error-laden Data Set 2. Here, the reproduction indices suggest anywhere between three sources and six sources. Clearly, the complications present in Data Set 2 (which are quite reasonable in terms of common environmental chemistry scenarios) are sufficient to introduce ambiguity between these various indices.

This ambiguity is due in part to the fact that all of the above methods are 'single-index' methods. Each involves calculation of a single numerical value or statistic, which represents the data set as a whole as a function of the number of principal components retained. The data analyst typically compares the behavior of the index as additional PCs are retained, relative to some rule-of-thumb cutoff criterion. The idea of a rule-of-thumb decision criterion (i.e., a minimum, a change in slope, a threshold) is troublesome in exploratory data analysis, because we have very little information to evaluate the efficacy of these rules. In such situations, we need other tools to gain deeper insight into the chemical system.

12.2.4.2 Variable-by-Variable Goodness of Fit

Miesch (1976) noted and addressed some of these problems. Miesch correctly observed that single index methods (in particular criteria based on percentage of variance) are misleading because they carry the tacit assumption that variability *not* accounted for by a reduced dimensional model is spread evenly across all originally measured variables. Miesch proposed instead, that goodness of fit be evaluated on a variable-by-variable basis. The variance accounted

Table 12.4

Reproduction indices for Data Set 1 (see Table 12.2).

PC No.	Eigenvalue	(1) Cumulative Percentage Variance	(2) Scree Test	(3) Normalized Varimax Loadings	(4) Malinowski Indicator Function	(5) Cross Validation PRESS(*i*)/ RSS(*i*−1)	(6) Signal to Noise Ratio
1	220.430	57.36	42.64	16	0.000674	NaN	39.32
2	135.150	92.54	7.46	8	0.000315	0.20	35.04
3	19.212	**97.53**	**2.47**	**16**	**0.000204**	**0.43**	**12.35**
4	1.198	97.85	2.15	1	0.000215	1.26	1.59
5	0.990	98.10	1.90	1	0.000229	1.41	1.35
6	0.861	98.33	1.67	0	0.000246	1.57	1.37
7	0.781	98.53	1.47	1	0.000266	1.76	1.41
8	0.719	98.72	1.28	1	0.000290	1.96	1.29
9	0.634	98.88	1.12	1	0.000318	2.19	1.34
10	0.556	99.03	0.97	0	0.000352	2.50	1.18
11	0.535	99.17	0.83	1	0.000392	2.79	1.27
12	0.475	99.29	0.71	1	0.000442	3.22	1.17
13	0.438	99.41	0.59	1	0.000503	3.67	1.21
14	0.382	99.50	0.50	1	0.000583	4.35	1.03
15	0.371	99.60	0.40	1	0.000681	5.00	1.15
16	0.317	99.68	0.32	1	0.000814	6.01	1.04
17	0.272	99.75	0.25	0	0.001001	7.11	1.09
18	0.213	99.81	0.19	1	0.001296	8.86	0.81
19	0.184	99.86	0.14	0	0.001769	11.31	0.81
20	0.169	99.90	0.10	0	0.002569	14.73	0.83
21	0.143	99.94	0.06	0	0.004157	21.04	0.84
22	0.126	99.97	0.03	1	0.007798	30.60	0.84
23	0.066	99.99	0.01	0	0.027834	308.49	0.64
24	0.043	100.00	0.00	0	NaN		0.00

for by *each* of the originally measured variables is evaluated for each potential number of principal components. Given an m sample by n variable data matrix $\mathbf{X}$ of rank m or n (whichever is smaller) the index used by Miesch (1976) was the 'coefficient of determination' (CD) between each variable in the original data matrix ($\mathbf{X}$), and its back-calculated reduced dimensional equivalent ($\hat{\mathbf{X}}$). For each number of potential eigenvectors, 1, 2 … *rank*, Miesch calculated an $n \times 1$ vector:

$$r_j^2 \cong \frac{s(x)_j^2 - (d_j)^2}{s(x)_j^2} \qquad (12.9)$$

where $s(x)_j^2$ is the variance of values in the *j*th column of $\mathbf{X}$, and $(d_j)^2$ is the variance of residuals between column j of $\mathbf{X}$ and column j of $\hat{\mathbf{X}}$. Miesch used the '≅' in this equation because he recognized that this was not a conventional r^2 or coefficient of determination ('CD') as defined by least squares linear regression. It is not the variance accounted for by the least

Table 12.5

Reproduction indices for Data Set 2 (see Table 12.3).

PC No.	Eigenvalue	(1) Cumulative Percentage Variance	(2) Scree Test	(3) Normalized Varimax Loadings	(4) Malinowski Indicator Function	(5) Cross Validation PRESS(*i*)/ RSS(i−1)	(6) Signal to Noise Ratio
1	585.600	77.92	22.08	50	0.000102	–	73.29
2	115.450	93.28	6.72	47	0.000059	0.33	31.89
3	17.807	**95.65**	4.35	**18**	0.000050	0.73	10.79
4	8.494	96.78	3.22	2	0.000046	**0.97**	6.54
5	4.751	97.41	2.59	1	0.000043	1.21	4.38
6	3.073	97.82	2.18	2	**0.0000421**	1.38	**3.30**
7	1.706	98.05	1.95	2	0.0000422	1.62	1.88
8	1.507	98.25	1.75	1	0.0000424	1.75	1.88
9	1.279	98.42	1.58	1	0.000043	1.88	1.85
10	1.055	98.56	1.44	0	0.000043	2.04	1.49
11	0.973	98.69	1.31	1	0.000044	2.19	1.60
12	0.898	98.81	1.19	1	0.000045	2.33	1.53
13	0.806	98.92	1.08	1	0.000046	2.54	1.51
14	0.782	99.02	0.98	1	0.000047	2.66	1.45
15	0.693	99.11	0.89	1	0.000048	2.81	1.31
16	0.596	99.19	0.81	1	0.000049	3.06	1.25
17	0.570	99.27	0.73	0	0.000050	3.29	1.28
18	0.532	99.34	0.66	1	0.000051	3.57	1.21
19	0.483	99.40	0.60	1	0.000053	3.83	1.16
20	0.434	99.46	0.54	1	0.000055	4.21	1.10
21	0.421	99.52	0.48	1	0.000056	4.52	1.14
22	0.371	99.57	0.43	0	0.000058	5.03	1.10
23	0.364	99.61	0.39	0	0.000060	5.25	1.10
24	0.313	99.66	0.34	0	0.000062	5.72	0.95
25	0.287	99.69	0.31	0	0.000065	6.33	0.94
26	0.274	99.73	0.27	0	0.000067	6.82	1.00
27	0.246	99.76	0.24	0	0.000070	7.74	0.93
28	0.245	99.80	0.20	0	0.000073	8.24	0.94
29	0.215	99.82	0.18	1	0.000076	8.89	0.88
30	0.178	99.85	0.15	0	0.000080	10.12	0.76
31	0.160	99.87	0.13	0	0.000084	11.25	0.72
32	0.142	99.89	0.11	1	0.000089	12.08	0.67
33	0.122	99.90	0.10	1	0.000095	13.78	0.62
34	0.111	99.92	0.08	0	0.000102	16.02	0.61
35	0.105	99.93	0.07	1	0.000109	18.24	0.58
36	0.093	99.95	0.05	0	0.000117	21.61	0.60
37	0.084	99.96	0.04	0	0.000125	22.87	0.54
38	0.060	99.96	0.04	0	0.000138	28.54	0.45
39	0.057	99.97	0.03	0	0.000152	32.49	0.44
40	0.047	99.98	0.02	0	0.000170	39.18	0.41
41	0.040	99.98	0.02	1	0.000191	46.38	0.40
42	0.031	99.99	0.01	0	0.000221	58.76	0.35
43	0.026	99.99	0.01	0	0.000260	73.10	0.32
44	0.020	99.99	0.01	0	0.000317	99.11	0.28
45	0.016	100.00	0.00	0	0.000396	140.67	0.26
46	0.014	100.00	0.00	0	0.000479	172.73	0.26
47	0.007	100.00	0.00	0	0.000697	317.28	0.19
48	0.003	100.00	0.00	0	0.001372	555.71	0.13
49	0.002	100.00	0.00	0	0.005275	–	0.12
50	0.002	100.00	0.00	0	–	–	0.00

squares regression line of $\mathbf{X}_j$ and $\hat{\mathbf{X}}_j$. Rather, the Miesch CD is the r^2 with respect to a line of one-to-one back-calculation between $\mathbf{X}_j$ and $\hat{\mathbf{X}}_j$. For CDs less than 1.0, the analyst must make some decision as to what value may be accepted. That decision is made in context of the analyst's experience, knowledge of measurement error (if available), and scientific context. As an example, if a certain PCB congener is known to be less accurate and precise using a certain gas chromatography (GC) column, the analyst may justifiably accept a lower CD for that congener, than for other congeners.

A graphical extension of Miesch's method has recently been implemented by Johnson, the CD scatter plot (Johnson, 1997; Johnson *et al.*, 2000). The appropriate graphic to illustrate the fit of the Miesch CD is a series of n scatter plots that show the measured value for each variable $\mathbf{X}_j$ plotted against the back-calculated values from the k proposed principal components ($\hat{\mathbf{X}}_j$). A scatter plot series for Data Set 2 is presented as Figure 12.5.

The scatter plot array shows 56 plots (one for each PCB congener) as back-calculated from a three-PC model. The Miesch CD is calculated and reported at the top left corner of each graph. When an insufficient number of principal components are retained there should be evident non-random deviations from the 1:1 fit line. For three principal components a good fit is observed for most congeners, but there is a systematic lack of fit observed for (1) non-detect censored data points (indicated as squares); (2) Sample 22 which had a data transcription error (triangle); and (3) the congener PCB 141, which coelutes with DDT. In particular, note that on many graphs, there are two 'non-detect' samples at high measured concentration. These are Samples 9 and 17 from Table 12.3. Both of these samples had low total PCB concentrations, and yielded non-detects for more than half of the reported PCB congeners.

The important point with regard to determining the number of significant principal components is that these errors, while not related to chemical sources fingerprints, are not random. Rather, they represent systematic signal within the data set, and thus they greatly influence indices shown in Tables 12.4–12.6.

The strength of CD scatter plots is that it allows rapid evaluation of (1) sample-by-sample goodness of fit, (2) variable-by-variable goodness of fit, and (3) outlier detection. Each of these is evaluated simultaneously, as a function of the number of principal components retained. That combination quickly leads the analyst to more insightful and direct questions than are possible based solely on what threshold numerical index one judges acceptable. The questions that we must now ask are:

1 What is the cause of deviation from good fit? Is it random error or systematic variability not accounted for by k PCs?

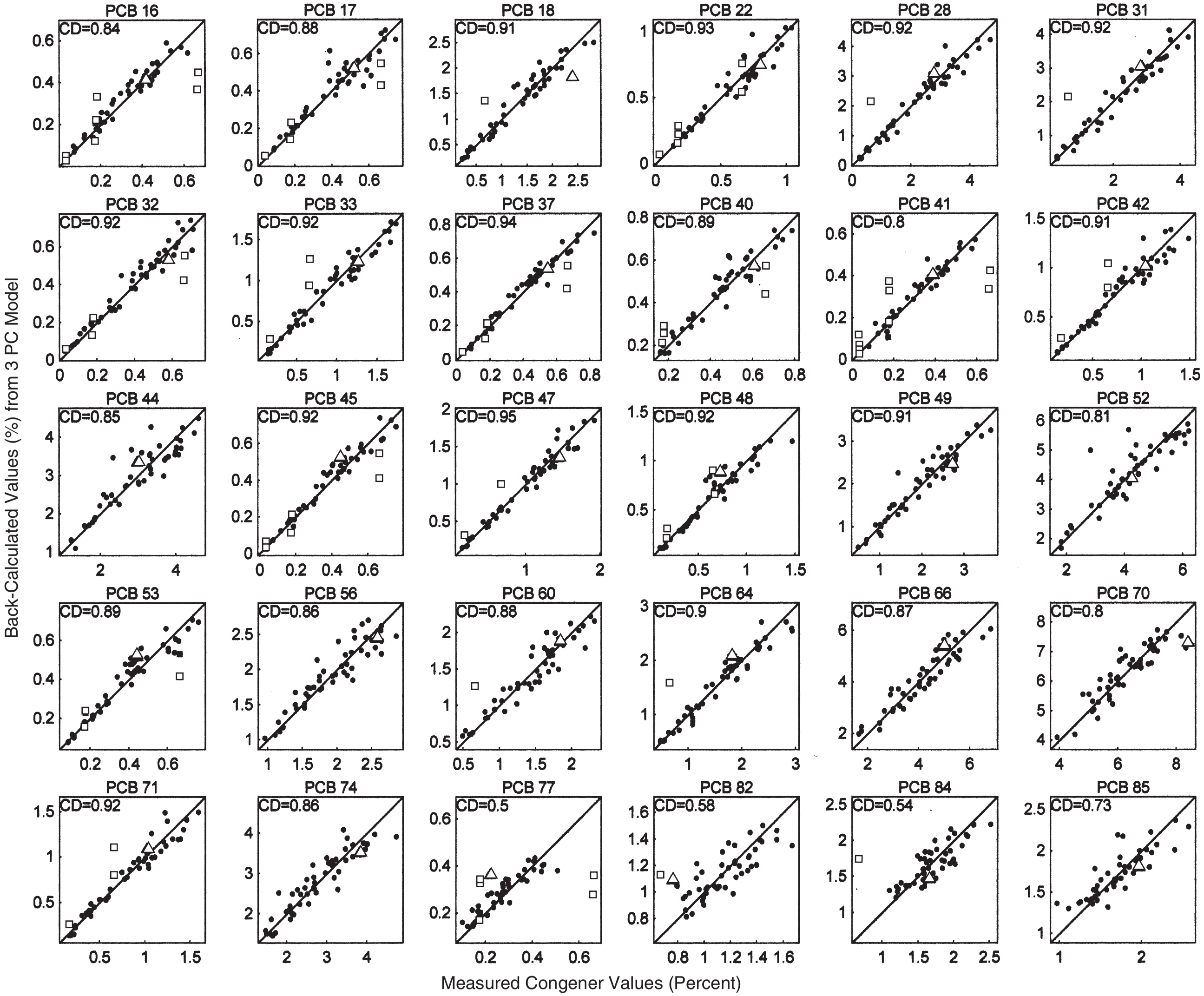


Figure 12.5a

Goodness-of-fit scatter plot array and Miesch coefficients of determination (CDs) for first 30 variables in Data Set 2. The x *axis is measured concentration. The* y *axis is the value back-calculated from a three principal component model. Non-detects (censored data points) are indicated as squares (□). Sample 22, which had a data transcription error, is indicated as a triangle (△).*

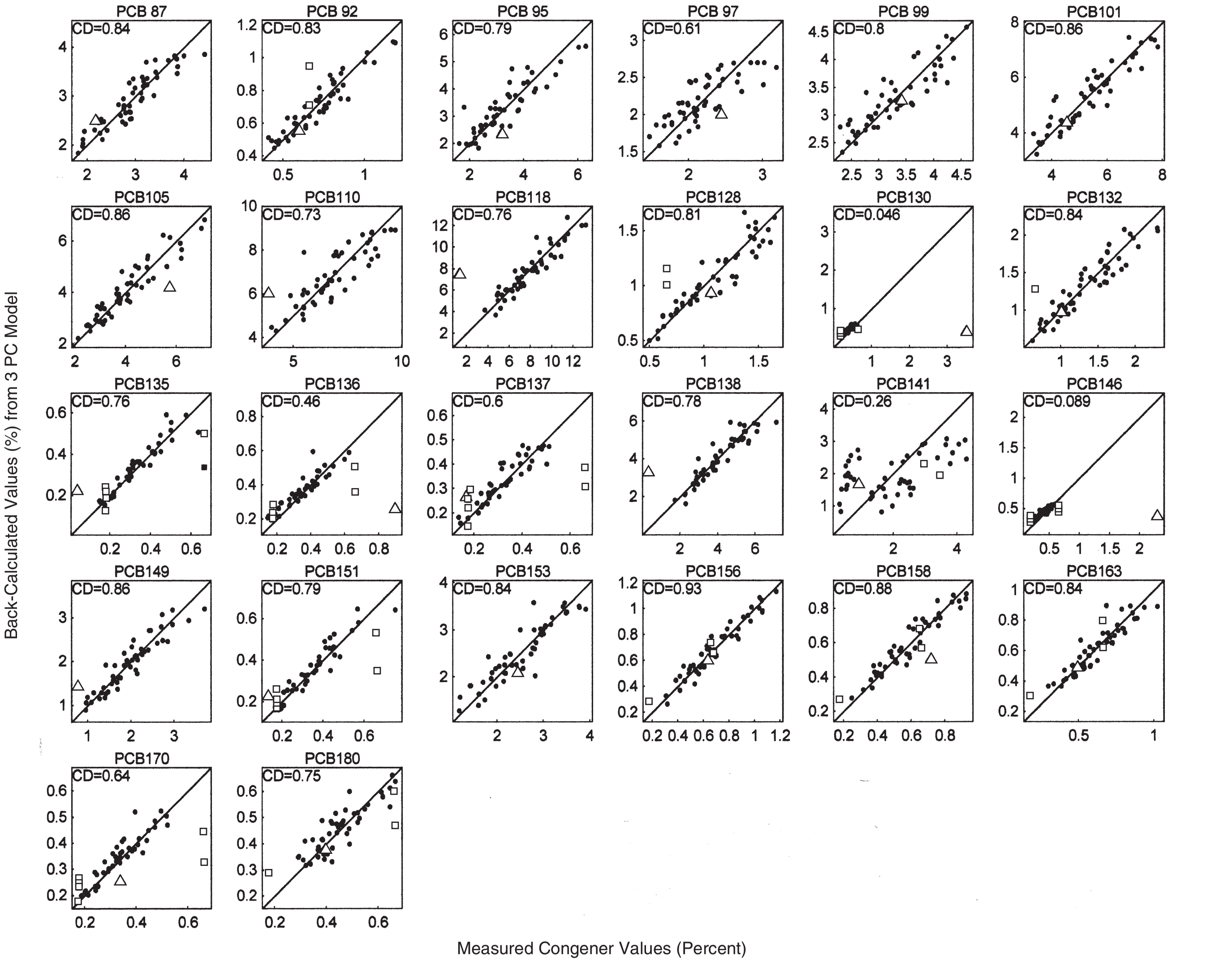


Figure 12.5b

Goodness-of-fit scatter plot array and Miesch coefficients of determination for last 26 variables in Data Set 2. The x *axis is measured concentration. The* y *axis is the value back-calculated from a three principal component model. Non-detects (censored data points) are indicated as squares (□). Sample 22, which had a data transcription error, is indicated as a triangle (△). Note lack of fit for PCB 141 due to DDT coelution.*

2 If the cause of the observed deviation is systematic, is it due to (1) data entry error, (2) analytical error, or (3) presence of an additional source of variability in the field?
3 How does one evaluate the number of significant PCs in the presence of such deviations?

In most cases, these questions cannot be answered in the realm of numerical data analysis. The analyst must now change hats and play the role of forensic scientist. The decision regarding how to deal with outlier samples must be considered in full context of the investigation: geographic/temporal distribution, analytical error, data entry error, method detection limits, as well as the possibility of an additional source, present in only one or a few samples. As discussed earlier, in environmental forensics investigations, decisions of 'significance' are often best made in the scientific context of the investigation, rather than through use of a rule-of-thumb numerical criteria. The use of such diagnostic plots is not new. They are standard in evaluation of linear regression models (Draper and Smith, 1981) but unfortunately are seldom used in evaluation of principal components models.

In the case of Data Set 2, the decision made by the data analyst is different for each type of outlier:

1 In the case of PCB 141 coeluting with DDT, the appropriate decision is to have the chemist go back and reanalyze the chromatograms to ensure that PCB 141 (a shoulder on the DDT peak) is correctly quantified.
2 In the case of the data transcription error, the appropriate decision is to correct the error in the spreadsheet, and rerun the analysis.
3 In the case of the two low concentration samples with multiple censored data points, usually the only realistic solution is to remove those two samples from the data set.
4 In the case of the remaining censored data points, those non-detects are generally at the low end of the measured range, and thus do not adversely affect the accuracy of back-calculation. We usually wish to retain as many samples as possible in the analysis. Therefore, in this case, we would typically leave these remaining non-detect samples in the matrix.

The changes above were made to Data Set 2. The PCA was rerun, and the revised reproduction indices are shown in Table 12.6. As for the much simpler Data Set 1, these indices are now in general agreement with each other, correctly indicating the presence of a three-component system.

12.2.5 PCA OUTPUT

As discussed in Section 12.2.1, the most common way of presenting PCA results is in terms of a PCA scores plot, where the analyst can evaluate relationships

Table 12.6

Reproduction indices for modified Data Set 2. For correctable errors (DDT coelution with PCB 141 and data transcription error in Sample 22) the data were modified accordingly. For uncorrectable problems (low concentration samples with many non-detects in two samples 9 and 17) were removed from the matrix.

PC No.	Eigenvalue	(1) Cumulative Percentage Variance	(2) Scree Test	(3) Normalized Varimax Loadings	(4) Malinowski Indicator Function	(5) Cross Validation PRESS(i)/RSS(i−1)	(6) Signal to Noise Ratio
1	561.330	76.26	23.74	46	0.000117	–	38.27
2	135.180	94.63	5.37	44	0.000059	0.25	14.83
3	19.919	**97.34**	**2.66**	**22**	0.000044	**0.57**	**2.55**
4	2.191	97.63	2.37	2	**0.0000434**	1.09	1.87
5	1.642	97.86	2.14	1	0.0000438	1.18	1.90
6	1.416	98.05	1.95	1	0.0000443	1.27	2.14
7	1.321	98.23	1.77	0	0.0000448	1.31	1.98
8	1.147	98.38	1.62	0	0.000046	1.40	1.76
9	1.025	98.52	1.48	1	0.000046	1.50	1.85
10	0.905	98.65	1.35	1	0.000047	1.61	1.59
11	0.839	98.76	1.24	1	0.000048	1.75	1.67
12	0.819	98.87	1.13	0	0.000050	1.77	1.58
13	0.712	98.97	1.03	0	0.000051	1.89	1.54
14	0.656	99.06	0.94	0	0.000052	2.02	1.54
15	0.612	99.14	0.86	1	0.000054	2.16	1.35
16	0.561	99.22	0.78	1	0.000055	2.31	1.37
17	0.511	99.29	0.71	1	0.000057	2.50	1.39
18	0.489	99.35	0.65	1	0.000059	2.74	1.30
19	0.482	99.42	0.58	1	0.000061	2.92	1.30
20	0.439	99.48	0.52	1	0.000063	3.13	1.31
21	0.412	99.53	0.47	0	0.000065	3.26	1.22
22	0.357	99.58	0.42	0	0.000068	3.60	1.16
23	0.339	99.63	0.37	1	0.000071	3.90	1.18
24	0.315	99.67	0.33	0	0.000074	4.15	1.07
25	0.276	99.71	0.29	0	0.000077	4.67	1.10
26	0.269	99.75	0.25	0	0.000081	4.83	0.94
27	0.227	99.78	0.22	0	0.000085	5.47	0.96
28	0.218	99.81	0.19	0	0.000089	6.03	0.89
29	0.199	99.83	0.17	1	0.000094	6.26	0.81
30	0.162	99.86	0.14	0	0.000101	7.05	0.77
31	0.148	99.88	0.12	1	0.000108	7.67	0.71
32	0.129	99.89	0.11	1	0.000116	8.79	0.71
33	0.119	99.91	0.09	1	0.000126	10.20	0.73
34	0.117	99.93	0.07	0	0.000136	11.61	0.68
35	0.106	99.94	0.06	0	0.000147	12.95	0.64
36	0.086	99.95	0.05	0	0.000162	15.57	0.59
37	0.079	99.96	0.04	0	0.000178	15.72	0.45
38	0.053	99.97	0.03	0	0.000204	19.88	0.46
39	0.050	99.98	0.02	0	0.000236	22.65	0.41
40	0.041	99.98	0.02	0	0.000279	28.45	0.40
41	0.037	99.99	0.01	0	0.000335	36.13	0.40
42	0.035	99.99	0.01	1	0.000403	40.03	0.31
43	0.020	99.99	0.01	0	0.000538	59.53	0.30
44	0.019	100.00	0.00	0	0.000740	74.99	0.26
45	0.013	100.00	0.00	0	0.001170	118.92	0.26
46	0.011	100.00	0.00	0	0.002074	175.26	0.19
47	0.006	100.00	0.00	0	0.006138	–	0.13
48	0.002	100.00	0.00	0	–	–	0.00

between samples on a two- or three-dimensional graphic. Scores plots for Data Set 1 were presented as Figures 12.2 and 12.3. For simple, clustered data sets such as this, the scores-plot visualization method is very effective. The problem, however, is that regardless of the complexity of the data or the results of goodness-of-fit diagnostics (Section 12.2.4) graphical limitations dictate a maximum of three principal components. This puts subtle pressure on the analyst to choose, whenever possible, three or fewer components. Therefore, the widespread occurrence of two and three component plots in the literature may be due more to this bias than to the inherent simplicity of environmental chemical systems. Another limitation to PCA scores plots is illustrated in Figure 12.6. This is the three-PC scores plot for cleaned-up Data Set 2. The corners of the gray shaded triangle represent the locations of the three Aroclor sources. All Data Set 2 samples plot within a triangle defined by these three vertices. This data cloud geometry is commonly observed when samples are mixtures of multiple sources.

Numerous software packages, including most general-purpose packages, perform PCA and allow the user numerous data transformation and visualization options. A few of these include Statistical Analysis System, (SAS: Cary, NC), Number Cruncher Statistical System (NCSS: Layton, Utah), Pirouette (Infometrix, Woodinville, WA).

As is evident in Figure 12.6, PCA can be effective in inferring the presence of a two- or three-component mixed system, but by itself PCA is not capable of determining the chemical compositions of the sources, or their relative contributions. In the case of analysis of mixtures, we need another technique in our numerical toolbox. Such tools are discussed in Section 12.3.

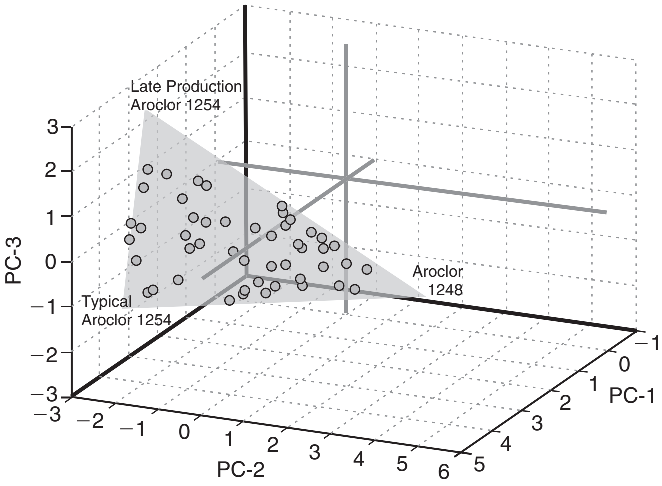


Figure 12.6

Three PC scores plot for Data Set 2, after errors have been corrected. No samples in Data Set 2 (dots) represent pure Aroclor sources. All are mixtures of the three sources. The approximate locations of pure sources are indicated at the vertices (corners) of the gray shaded triangle.

12.3 SELF-TRAINING RECEPTOR MODELING METHODS

An increasingly common approach in environmental forensic investigations involves the use of receptor models. These models are designed to resolve three parameters of concern in a multivariate, mixed chemical system: (1) the number of components in the mixture, (2) the identity (i.e., chemical composition or fingerprints) of each component, and (3) the relative proportions of each component in each sample. These objectives are stated mathematically as the determination of k, **A** and **F** in Equation 12.1. The source apportionment equation is similar to the scores and loadings expression given in Equations 12.5 and 12.6. However, principal component scores and loadings are abstract, orthogonal matrices, and do not represent feasible chemical compositions or source contributions. They are abstract in that when expressed in terms of the original chemical variables, principal component loadings typically contain negative elements.

Receptor modeling methods therefore involve rotation of matrices **A** and **F** to an oblique solution derived within reduced (k-dimensional) principal component space. The rotation is performed per explicit non-negativity constraints imposed on matrices **A** and **F**. For example, a fingerprint cannot have a chemical composition with −10% PCB 138, thus matrix **F** cannot have negative elements. Similarly, a sample cannot have a −35% contribution from a given source, thus matrix **A** cannot have negative elements (Miesch, 1976a; Full *et al.*, 1981, 1982; Gemperline, 1984; Hopke, 1989; Henry and Kim, 1990; Kim and Henry, 1999). Non-negativity constraints are typical requirements for all mixing models. In addition, other explicit restrictions may be imposed, if one has additional knowledge of physical/chemical constraints on the system (Henry and Kim, 1990; Johnson, 1997; Kim and Henry, 1999).

Multivariate receptor modeling methods use PCA as an intermediate step to determine the number of contributing sources, and to provide a reduced dimensional reference space for resolution of the model. In all of the receptor modeling methods discussed here, determining the number of sources (k) essentially reduces to the problem of choosing the number of significant principal components. As such, discussions regarding determination of significant principal components (Section 12.2.4) are equally relevant to receptor modeling.

An assumption of the conceptual mixing model/receptor model is that the system must be over-determined. That is, the data set must contain more variables or samples (whichever number is smaller) than there are sources. If we measure only four chemicals in each sample, and six sources contribute to the contamination, we cannot completely or realistically resolve the model. Another assumption is that of linear mixing. We assume that the relative proportions of variables in each source are fixed, and that source contributions are linearly additive. That is, as we increase the proportion of a source fingerprint in a sample, the variables that are characteristic of that fingerprint will increase proportionally (linearly) in that sample.

After the choice of k, (Section 12.2.4) the receptor model then resolves the chemical compositions of sources (**F**) and the contributions of the sources in each of the samples (**A**). Recall, however, that in environmental forensics investigations, we rarely have such *a priori* knowledge of all sources. If possible, we would like to derive source patterns directly from analysis of ambient data. Three such methods have been used in environmental source apportionment investigations: (1) the DENEG algorithm used in polytopic vector analysis (PVA) (Full *et al.*, 1981, 1982); (2) the unique vector iteration method used in target transformation factor analysis (TTFA: Roscoe and Hopke, 1981; Gemperline, 1984; Hopke, 1989; Malinowski, 1991); and (3) source apportionment by factors with explicit restrictions (SAFER) method, used in extended self-modeling curve resolution (ESMCR: Henry and Kim, 1990; Henry *et al.*, 1994; Kim and Henry, 1999).

These three methods are analogous in that (1) they do not require a training data set; (2) they are PCA based methods; (3) they involve solution of quantitative source apportionment equations by development of oblique solutions in PCA space; and (4) each involves the use of non-negative constraints.

The full PVA algorithm has not been set out in any single paper. This chapter provides the opportunity to do so. As such, the mathematics of PVA will be discussed in greater detail than the other two methods. TTFA and ESMCR are presented to provide the reader with an intuitive understanding of how those algorithms operate. The reader is referred to the original papers for specifics of those algorithms. Each of these three methods were applied to the PCB data set given in Table 12.2, and each yielded source compositions that closely matched the compositions of the Aroclor sources in Table 12.1. While we will focus on these three methods, it should be noted that there are yet other methods with similar objectives, which have been described in the literature (Ozeki *et al.*, 1995; Xie *et al.*, 1998).

12.3.1 POLYTOPIC VECTOR ANALYSIS (PVA)

PVA was developed for analysis of mixtures in the geological sciences, but it has evolved over a period of 40 years, with different aspects of the algorithm presented in a series of publications, by a number of different authors. The roots of PVA are in principal components analysis, pattern recognition, linear algebra, and mathematical geology. Its development in mathematical geology can be traced back to the early 1960s and John Imbrie, a paleontologist

(Imbrie, 1963). Following Imbrie's work, a series of FORTRAN-IV programs were published (Manson and Imbrie, 1964; Klovan, 1968; Klovan and Imbrie, 1971). The resulting software developed by Imbrie (at Brown University) and Ed Klovan (at the University of Calgary) was called CABFAC (*C*algary and *B*rown *FAC*tor Analysis) and quickly became the most commonly used multivariate analysis algorithm in the geosciences. Subsequent investigators that proved crucial in development of the PVA algorithm included A.T. Miesch and William Full. Miesch, a geochemist with the US Geological survey in the 1970s, was one of the first to take full advantage of Imbrie's oblique vector rotation methods (Miesch 1976a,b). Miesch also developed the variable-by-variable goodness-of-fit criteria (Miesch CDs) discussed in Section 12.2.4.2. William Full, as a PhD candidate at the University of South Carolina in the early 1980s, developed the DENEG algorithm, which allows end-members (sources) to be resolved without *a priori* knowledge of their composition, and without use of a training data set (Full *et al.*, 1981, 1982).

The name 'polytopic vector analysis' follows directly from the jargon of Imbrie (1963) and Full *et al.* (1981, 1982). Imbrie (and many others) referred to eigenvector decomposition models, resolved in terms of orthogonal axes as 'factor analysis'. Solutions resolved in terms of oblique vectors he termed 'vector analysis'. Imbrie's is not the definition of true factor analysis as defined by Harman (1960). Regardless, Imbrie's terminology has held within mathematical geology and a number of other subdisciplines. Because PVA involves resolution of oblique vectors as source compositions, thus the term vector analysis. The term 'polytopic' is due to the fact that PVA involves resolution of a $k - 1$ dimensional solid, a 'simplex' or 'polytope', within k dimensional principal component space. As illustrated in Figure 12.6, the polytope or simplex in three-dimensional space is a two-dimensional triangle. In two space it is a straight line. In four space it is a tetrahedron. Because the objective of PVA is resolve a $k - 1$ dimensional simplex, the constant vector length transform is typically employed because it forces all sample vectors to have unit length, and thus all data are constrained to a $k - 1$ dimensional space within k space (see Figure 12.4 – Section 12.2.2).

This section presents the full PVA algorithm running under default conditions; i.e., (1) the EXTENDED CABFAC algorithm (Klovan and Imbrie, 1971) and (2) the iterative oblique vector rotation algorithm originally presented as EXTENDED QMODEL (Full *et al.*, 1981, 1982). While any number of alternative data transformation and calculation options are available and may be implemented, these algorithms represent the core of PVA as it is presently implemented under default options by the commercial version of the SAWVECA (*S*outh *C*arolina *a*nd *W*ichita *Ve*ctor *A*nalysis: Residuum Energy, Inc., Dickinson, TX). SAWVECA performs the dimensionality analysis of the

Klovan and Miesch's EXTENDED CABFAC; along with the CD scatter plot goodness-of-fit diagnostics of Johnson (1997; Johnson *et al.*, 2000). SAWVECB is the EXTENDED QMODEL and FUZZY QMODEL algorithms of Full *et al.* (1981, 1982). Readers not interested in the formalism of the PVA algorithm can skip ahead to Section 12.3.1.4.

12.3.1.1 Scaling Functions: Back-Calculation to Original Metric

The transformations presented in Section 2.2.2 serve to optimize the eigenvector decomposition, but interpretation and evaluation of matrices **A**″ and **F**″ in any scientific context is difficult. Calculations are best performed in transform metric, but evaluation/interpretation must be done in measurement metric. Thus the results must be back-calculated. Where double prime (e.g. **X**″) indicates a matrix which has been transformed in turn by the range transform and the constant vector length transform, the mapping functions presented by Miesch (1976a) allow us to translate the equations $\hat{\mathbf{X}}'' = \mathbf{A}''\,\mathbf{F}''$ back to $\hat{\mathbf{X}} = \mathbf{AF}$ (percent or 'constant row-sum' metric). The mathematics are discussed below. As scaling is called upon in various portions of PVA, the reader is referred back to this section for a description. The first step in back-calculation to measurement metric is the definition of what Miesch termed scale factors: s_k. Given k retained eigenvectors, Miesch defines a $1 \times k$ row vector of scale factors $s = \{s_k\}$ where each element s_k is defined as:

$$s_k = \frac{\mathrm{K} - \sum_{j=1}^{n} x_{\min j}}{\sum_{j=1}^{n} (f'_{kj}\,(x_{\max j} - x_{\min j}))} \tag{12.10}$$

K is a constant (the sum of each row: usually 100), f''_{kj} is the element of the scores matrix, and $x_{\max j}$ and $x_{\min j}$ are the maximum and minimum values in the jth column of the original data matrix **X**.

The elements of the back-calculated scores matrices **F**′ and **F** are then calculated, in turn, as follows:

$$f'_{kj} = s_k\, f''_{kj} \tag{12.11}$$

$$f_{kj} = f'_{kj}\,(x_{\max j} - x_{\min j}) + x_{\min j} \tag{12.12}$$

Similarly, the elements of the back-calculated loadings matrices **A**′ and **A** are calculated as:

$$a'_{ik} = \frac{a''_{ik}}{s_k} \tag{12.13}$$

If **r** is then defined as a column vector of the *m* row-sums of **A**′, the elements of **A** are:

$$a_{ik} = \frac{a'_{ik}}{r_i} \tag{12.14}$$

12.3.1.2 Eigenvector Decomposition and Determining the Number of Sources

A singular value decomposition is carried out on transformed matrix **X**′ or **X**″ as presented in Equation 12.7. The results are translated into scores and loading ($\mathbf{A}''_Q$ and $\mathbf{F}''_Q$) again, as per Equation 12.7. The scores and loadings matrices are then translated back to constant sum metric (**A** and **F**) using the scaling functions presented in the preceding section (Equations 12.10 and 12.14). The task then is determination of *k*, the appropriate number of sources. As discussed in Section 12.2.4, this reduces the problem of choosing the number of significant principal components, and all discussions presented in that section are equally relevant here. The methods used most often in PVA are (1) the normalized loading criteria (Ehrlich and Full, 1987) and (2) the Miesch CDs and scatter plots as described in Section 12.2.4.2. For each number of potential principal components ($k = 1, 2, \ldots n$, if n is smaller; or $k = 1, 2, \ldots m$, if m is smaller), $\hat{\mathbf{X}}$ is calculated as:

Matrix dimensions
$$\underset{(m\times n)}{\hat{\mathbf{X}}} = \underset{(m\times k)}{\mathbf{A}}\ \underset{(k\times n)}{\mathbf{F}} \tag{12.15}$$

The Miesch CDs are calculated and scatter plots constructed using the reduced dimensional scores and loadings, as expressed in percentage metric.

12.3.1.3 Determining End-Member Compositions and Mixing Proportions

Following determination of *k* (the number of end-members) a set of mathematical procedures are used to resolve the second and third objectives of the receptor modeling problem: (2) determine the composition of the end-members, and (3) determine the relative proportions of each end-member in each sample. This process is termed polytope resolution. The polytope resolution phase is typically conducted within *k* dimensional varimax space, but can be equivalently performed in unrotated principal component space. The first *k* columns of **A**″ are taken, yielding a reduced matrix **A**″ of size $m \times k$. Because $\hat{\mathbf{X}}'' = \mathbf{A}''\ \mathbf{F}''$, we may determine matrix **F**″ by the following matrix regression equation:

$$\mathbf{F}'' = (\mathbf{A}''^{T}\ \mathbf{A}'')^{-1}\,\mathbf{A}''^{T}\ \mathbf{X}'' \tag{12.16}$$

Using the scaling equations in Section 12.3.1.1, matrices **A**″ and **F**″ are then transformed back to the original constant row-sum metric, and the estimate of $\hat{\mathbf{X}}$ is calculated as:

Matrix dimensions
$$\underset{(m\times n)}{\hat{\mathbf{X}}} = \underset{(m\times k)}{\mathbf{A}}\ \underset{(k\times n)}{\mathbf{F}} \qquad (12.17)$$

12.3.1.3.1 Selection of Initial Polytope

The first task of the polytope phase is an initial estimate of a polytope. A number of techniques have been proposed. Most commonly employed is the 'extended' method, so named because it was used by Full *et al.* (1981) in the EXTENDED QMODEL Fortran IV algorithm. The formalism of the extended method is described below. Alternative methods, their advantages and disadvantages are subsequently discussed.

The extended method establishes the initial polytope by taking the k most mutually extreme samples in the data set as vertices (the EXRAWC subroutine of Klovan and Miesch, 1976). EXRAWC first picks a good candidate set for these k samples: those with maximum loadings on each of the first k factors of **A**″. The columns of **A**″ are scanned, and the maximum absolute value loadings in each column are identified. The rows (samples) of **A**″ corresponding to the maximum loadings are then put into a new $k \times k$ matrix $\mathbf{O}_0$. This operation is done without duplication (i.e. no two rows of $\mathbf{O}_0$ are the same). Clearly, the samples that make up $\mathbf{O}_0$ are candidates for the k most mutually extreme samples. The PVA algorithm uses these k vectors as oblique reference axes for all samples. The resultant oblique loadings and scores matrices $\mathbf{A}_0''$ and $\mathbf{F}_0''$ are calculated as:

$$\mathbf{F}_0'' = \mathbf{O}_0\, \mathbf{F}'' \qquad (12.18)$$

$$\mathbf{A}_0'' = \mathbf{A}''\mathbf{O}_0^{-1} \qquad (12.19)$$

Matrices $\mathbf{F}_0''$ and $\mathbf{A}_0''$ are then scaled back to measurement space using the scaling functions described in Section 2.2.4, yielding $\mathbf{F}_0$ and $\mathbf{A}_0$.

Matrix $\mathbf{A}_0$ is then inspected to determine if the k samples in $\mathbf{O}_0$ are indeed the k most mutually extreme. Following the method of Imbrie (1963) and Miesch (1976a), if the maximum loadings in matrix $\mathbf{A}_0$ equal 1.0 and correspond to the samples taken for matrix $\mathbf{O}_0$, then the k most mutually extreme samples have been taken and matrix $\mathbf{O}_0$ is then used as initial oblique reference axis for iterative model resolution. If loadings greater than 1.0 are identified in $\mathbf{A}_0$, the sample(s) corresponding to that maximum loading replaces the original sample in $\mathbf{O}_0$, and the process is repeated until a suitable set of k samples is obtained. If the algorithm finds no set of samples with all elements less than 1.0, then the original matrix $\mathbf{O}_0$ is taken as the initial set of oblique vectors. The algorithm for determining extreme samples is termed the EXRAWC procedure.

Outlier samples related to analytical problems or human error should be corrected or omitted from the analysis, as discussed in Section 12.2.4. However, if such errors are not corrected, the extended method can perform poorly. Outliers are extreme samples, and the extended method defines the initial polytope using extreme samples as vertices. In such cases, another method based on the fuzzy *c* means clustering algorithm of Bezdek (Bezdek, 1981; Bezdek *et al.*, 1984) will often produce better results. The mathematics of so-called FUZZY QMODEL are presented by Full *et al.* (1982) demonstrated that the *k* fuzzy cluster centers as defined in *k* dimensional eigenspace were always well within the convex hull defined by the sample data cloud, and were minimally affected by the presence of outliers. The vectors that define the *k* fuzzy cluster centers are taken as the row vectors of $\mathbf{O}_0$, and matrices $\mathbf{A}_0$ and $\mathbf{F}_0$ are defined relative to these vectors, as discussed above. The main disadvantage of the fuzzy method is that it is more computationally expensive to run. In the absence of outliers due to error (as we hope would result from diligent outlier identification in the PCA step) no advantage is gained by choosing fuzzy over extended.

A third option is to use the samples with maximum loadings in $\mathbf{A}''$. In instances where the EXRAWC subroutine does not converge, the EXTENDED option will default to the set of samples with maximum loadings.

A final option is external input of end-member compositions. If end-member compositions are known or suspected prior to the analysis, those suspected sample compositions may be plugged into the model as potential end-members. Suppose for instance that Aroclors 1248, and the two 1254 variants *were* suspected as the contributing sources for Data Set 2. If those source compositions were used as external end-members, and the model converged without iteration, the tested end-members would be considered feasible. This method is essentially the same as target testing as described by Hopke (1989) and Malinowski (1991) and is also similar to methods that require use of a training data set (such as chemical mass balance approaches). While this option is included in PVA software, in practice it is seldom used, because it constitutes a tacit hypothesis test, which is contrary to an exploratory data analysis philosophy. Regardless of the method used, the result is ultimately a set of *k* vectors as the rows of matrix $\mathbf{O}_0$. Matrices $\mathbf{A}_0$ and $\mathbf{F}_0$ are then defined relative to these vectors.

12.3.1.3.2 Testing Matrices $\mathbf{A}_0$ and $\mathbf{F}_0$ for Negative Values

Once an initial polytope is defined, the algorithm scans matrices $\mathbf{A}_0$ and $\mathbf{F}_0$ for negative values. The DENEG subroutine described below distinguishes between adjustable negative values and nonadjustable negative values, based on three user-defined numerical criteria. The first, t_1, is the 'mixing proportions cutoff criterion'. The default t_1 is -0.05. In other words, the algorithm will allow up to a -5% mixing proportion in any sample. The purpose of t_1 is to allow for some noise in the model.

If $\mathbf{A_0}$ does contain negative matrix elements less than t_1, the matrix is again scanned using a second criterion, $t_{2,}$ referred to as the DENEG value. The default DENEG value is -0.25, but may be modified as the user sees fit. The DENEG subroutine recognizes adjustable mixing proportions only if they fall in the range between t_1 and t_2. At first glance, the need for the t_2 is not obvious, but in effect, it serves as our final line of defense against outliers. In early development of the algorithm, Full observed that in the presence of outliers, a model would often converge (i.e., no negatives less than -0.05) with the lone exception of an outlier with an extremely high negative value (less than -25%) in matrix **A**. By using the t_2 criterion, the DENEG subroutine would not iterate in an attempt to fit that single sample into the model.

Finally, the algorithm scans $\mathbf{F_0}$ for negative values using a third criteria, t_3, referred to as the 'end-member composition cutoff criterion'. Again, the default t_3 value is -5%, and serves the purpose of allowing some noise in the model. If there are no adjustable negative elements in matrices $\mathbf{A_0}$ and $\mathbf{F_0}$, the algorithm stops. $\mathbf{A_0}$ is taken as the mixing proportions matrix and $\mathbf{F_0}$ as the end-member compositions matrix.

12.3.1.3.3 The DENEG Algorithm

If $\mathbf{A_0}$ and $\mathbf{F_0}$ contain adjustable negatives, the program starts an iterative process of expansion and rotation of the polytope until two criteria are met: (1) mixing proportions have no adjustable negative values, and (2) end-member compositions have no adjustable negative values.

Geometrically the DENEG procedure is a process of alternate polytope expansion and rotation. The results of DENEG applied to the synthetic three-source PCB data set are shown graphically in Figure 12.7. Recall that data transformations are performed such that each sample vector has unit length (be that in measurement space or reduced dimensional principal components space). Thus, all data points plot on sphere, unit length from the origin (Figure 12.7). This feature of the data holds in reduced dimensional space; thus, all data vectors in three-dimensional PC space would also have unit length (Figure 12.7). DENEG begins by constructing an initial simplex in principal component space (Iteration 0: Figure 12.7). The EXTENDED method (Section 12.3.1.3.1) was used to define the initial polytope (Iteration 0), so the vertices of the Iteration 0 triangle are located at the three most mutually extreme samples in the data set. Had pure end-members been contained within the data set, the non-negativity criteria would have been met at Iteration

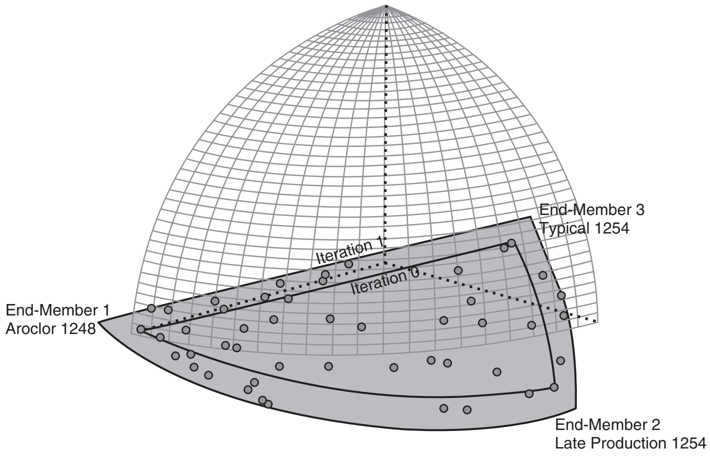


Figure 12.7

Graphical representation of DENEG algorithm applied to Data Set 2, after errors have been corrected. In three-PC space, the algorithm establishes a initial simplex with vertices at the three most mutually extreme samples in the data set (Iteration 0), then alternately expands and rotates the simplex until non-negativity constraints are met. The algorithm converged at Iteration 1.

0, and the algorithm would have converged without iteration. However, this was not the case. Negatives are present in $\mathbf{A_0}$ (i.e., samples are located outside the Iteration 0 simplex: Figure 12.7). Thus DENEG begins a series of iterations which alternately adjusts **A** and **F** until neither matrix contains negative values. The DENEG algorithm converged after one iteration, and the Iteration 1 simplex is shown in Figure 12.7 as the shaded triangle. The associated matrix operations are described below.

We begin by defining a 1 × k row vector **D** = whose elements d_i ($i = 1, 2, \ldots k$) are the lowest adjustable mixing proportions in columns 1, 2, … k of matrix $\mathbf{A_0}$. We also define a scalar z:

$$z = \frac{1}{(1 + d_1 + d_2 + \cdots + d_n)} \tag{12.20}$$

A new matrix $\mathbf{A_1} = \{a_{1_{ij}}\}$ may then be calculated as:

$$a_{1_{ij}} = (a_{0_{ij}} + d_j) \times z \tag{12.21}$$

The resultant matrix $\mathbf{A_1}$ represents an end-member mixing proportions matrix, with no samples composed of negative mixing proportions. Geometrically, Equation 12.21 has the effect of moving the edges of the polytope out in an edge parallel direction, and stopping at the outermost sample from that edge: an edge parallel expansion. The corresponding matrix $\mathbf{F_1}$ is then calculated by matrix regression as

$$\mathbf{F_1} = (\mathbf{A_1}^T \mathbf{A_1})^{-1} \mathbf{A_1}^T \hat{\mathbf{X}} \tag{12.22}$$

Matrix $\mathbf{F_1}$ is the end-member compositions matrix. The algorithm scans $\mathbf{F_1}$ for adjustable negatives (less than t_3). If none are encountered, the algorithm ceases iterating, and $\mathbf{F_1}$ and $\mathbf{A_1}$ are the final mixing proportions and end-member composition and mixing proportions matrices. If adjustable negatives are encountered in $\mathbf{F_1}$, the algorithm continues to the polytope rotation part of the algorithm.

If $\mathbf{F_1}$ contains adjustable negatives, the algorithm continues by substituting all negative values of $\mathbf{F_1}$ with zeros. A new matrix $\mathbf{F_2}$ is then defined by renormalizing the rows of the non-negative $\mathbf{F_1}$ to sum to 100%. The algorithm scales $\mathbf{F_2}$ using the transforms described in a previous section. Matrix $\mathbf{F_2''}$ is then taken as a prospective end-member compositions matrix. $\mathbf{F_2}$ is scaled to transform space, and is projected down into k dimensional space as a new set of factor loadings: matrix $\mathbf{O_2}$. Oblique loadings matrix is then defined as:

$$\mathbf{F_2''} = \mathbf{O_2}\ \mathbf{F''} \tag{12.23}$$

$$\mathbf{A_2''} = \mathbf{A''}\ \mathbf{O_2}^{-1} \tag{12.24}$$

We then scale matrices $\mathbf{A_2''}$ back to measurement space using the scaling functions described in Section 12.3.1.1 , yielding $\mathbf{A_2}$. Matrix $\mathbf{A_2}$ is then inspected for adjustable negatives. If none are encountered, the program ceases iterating. If adjustable negatives are encountered in $\mathbf{A_2}$, the iterations continue. The algorithm redefines matrices $\mathbf{A_2}$ and $\mathbf{F_2}$ as $\mathbf{A_0}$ and $\mathbf{F_0}$ and DENEG loops back to the beginning. In the event that iterations do not result in non-negative matrices, two additional cut-off criteria are defined. Criterion t_4 is a measure of how similar one iteration is to the next. Criterion t_5 is the user-defined maximum number of iterations.

12.3.1.4 Results of PVA Applied to Data Set 2

PVA was run on Data Set 2, modified as indicated in Section 12.2.4.2. Using Miesch's scaling functions (Section 12.3.1.1) the vertices of the Iteration 1 triangle (Figure 12.7) may be back-calculated to percentage metric, to yield estimates of source compositions (matrix **F**). The mixing proportions matrix (**A**) is also determined by back-calculation to the original percentage metric. The resolved end-member compositions are illustrated in Figure 12.8. The three patterns resolved by PVA are clearly quite similar to the true Aroclor compositions.

PVA has been applied in a number of multivariate chemical source apportionment investigations, primarily in sediment and water studies (Ehrlich *et al.*, 1994; Doré *et al.*, 1996; Jarman *et al.*, 1997; Bright *et al.*, 1999; Johnson *et al.*, 2000).

12.3.2 Unique Vector Rotation Method

PVA is one 'self-training' method that allows source profiles to be derived in absence of *a priori* knowledge of their chemical composition, but other such

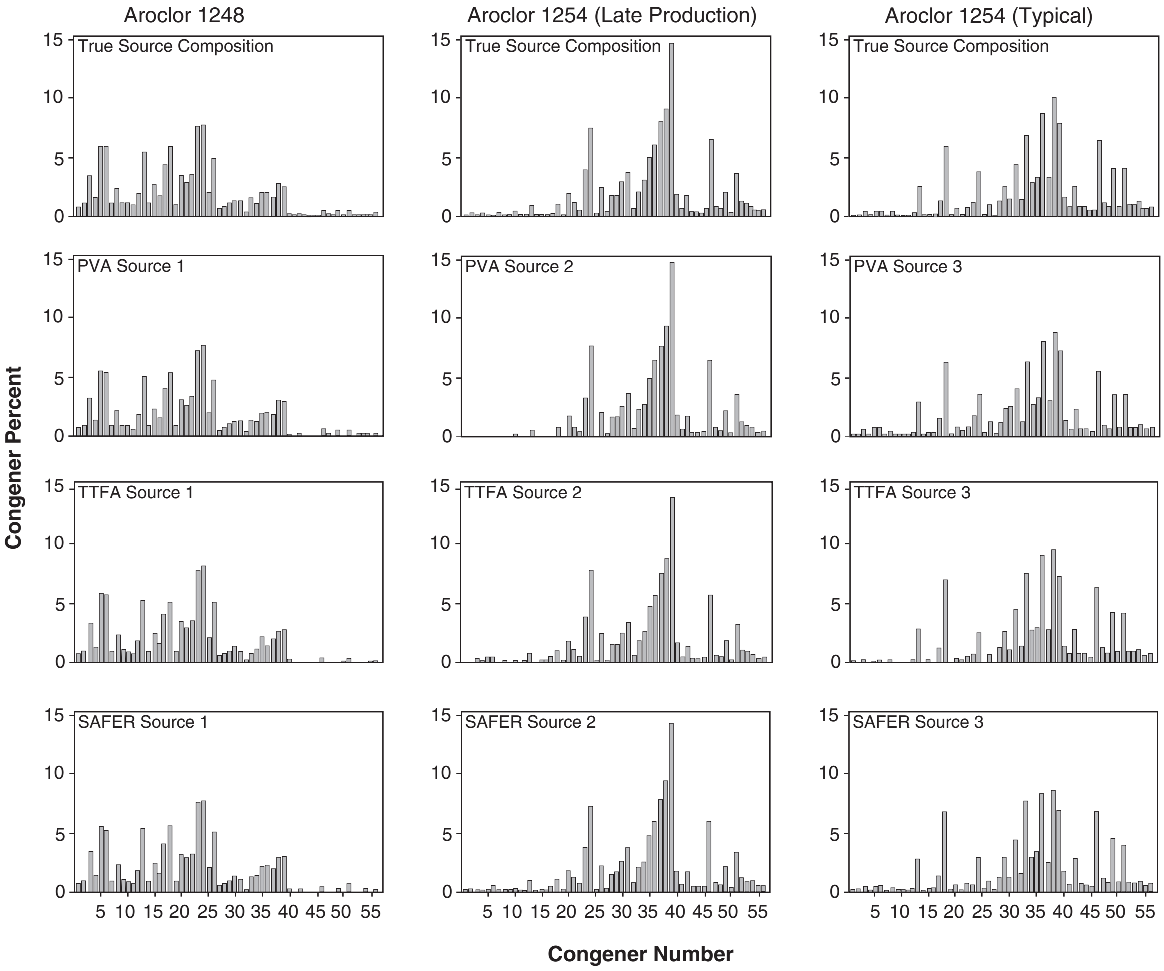


Figure 12.8

Comparison of true Aroclor compositions (top row) with source compositions derived through PVA (row 2), TTFA (row 3) and SAFER (row 4). Congener numbers correspond to congener order presented in Table 12.2.

methods have seen considerable application to environmental chemical data. One of these is target transformation factor analysis (TTFA), which developed within analytical chemistry/chemometrics rather than mathematical geology/geochemistry (Roscoe and Hopke, 1981; Gemperline, 1984; Hopke, 1989; Malinowski, 1991).

In TTFA, the subroutine that allows estimates of source composition in the absence of known sources is the unique vector rotation method (Roscoe and Hopke, 1981; Gemperline, 1984; Hopke, 1989; Malinowski, 1991). This method begins by establishing a $n \times n$ matrix where each row vector is 100% of a single analyte (i.e. 'unique vectors'). In turn, each of these vectors is iteratively rotated within principal component space. For Data Set 2, Figure 12.9 shows the rotation trajectories for each of the 56 unique vectors. Like DENEG, the method often involves transformation of sample vectors to unit length

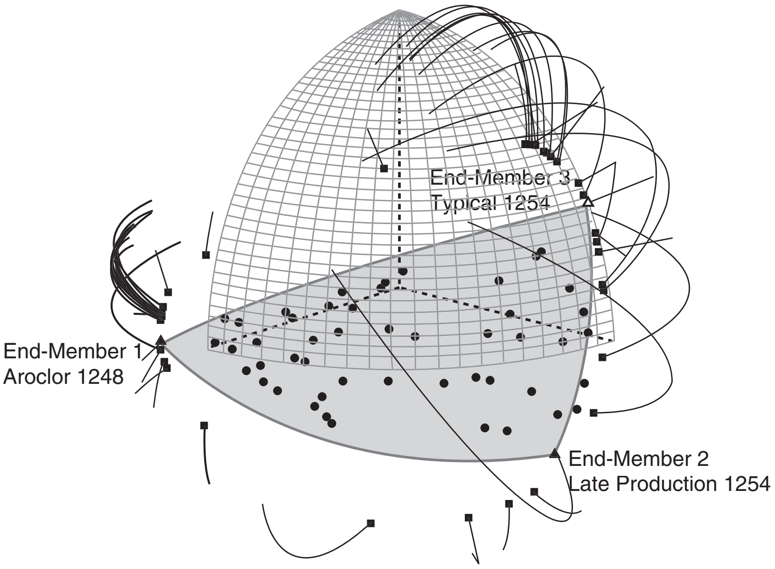


Figure 12.9

Graphical representation of the unique vector iteration method as applied to Data Set 2, after errors have been corrected. This method is used in target transformation factor analysis and involves iterative rotation of a series of n unique vectors.

(Hopke, 1989). As such, Figure 12.9 illustrates the iterative rotations within the constant vector length metric. Each of the iteration trajectories is shown as a curved line on Figure 12.9, which terminates at a location marked (■). The 56 resultant oblique vectors represent a matrix of candidate source profiles for the receptor model. Three of the candidate profiles (▲) were chosen as source compositions in this example, and the shaded simplex shown is the ternary diagram (simplex) constructed based on the chosen profiles.

The unique vector iteration method has proven useful in application in a number of source apportionment investigations, particularly in air receptor modeling (Chang *et al.*,1988; Hopke, 1989; Borbély-Kiss *et al.*, 1993; Moro *et al.*, 1997). However, the method can be extremely cumbersome, especially in a situation such as Data Set 2, where a large number of variables (n = 56) necessitates evaluation of the feasibility of many permutations of candidate source profiles. The TTFA software package FANTASIA (Factor Analysis to Apportion Sources in Aeresols) is available through Dr Philip Hopke, Clarkson University, Potsdam, New York. Software for Positive Matrix Factorization (referenced herein, but not discussed in detail) may also be obtained through Dr Hopke.

12.3.3 SAFER Method

Another receptor modeling method, SAFER (Source Apportionment by Factors with Explicit Restrictions) is used in extended self-modeling curve resolution (ESMCR: Henry and Kim, 1990; Kim and Henry, 1999). Unlike PVA and TTFA, ESMCR does not typically involve transformation to unit length. As

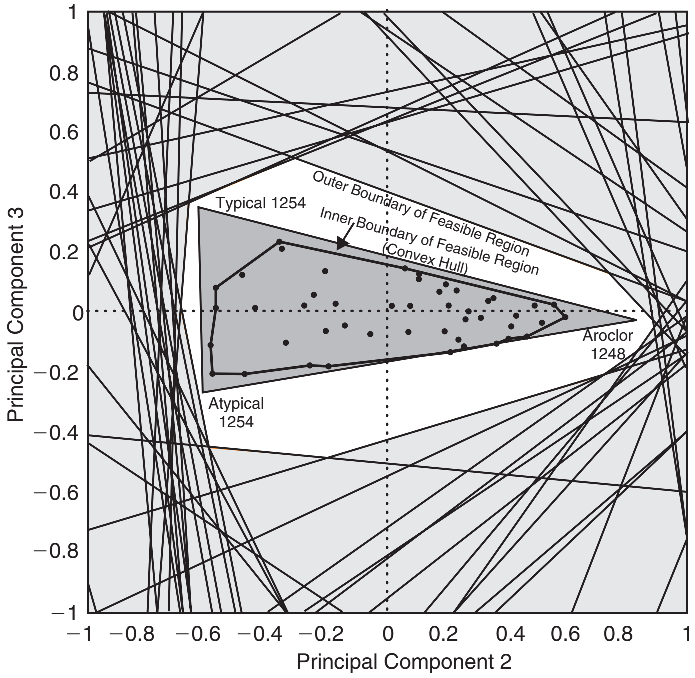


Figure 12.10

Graphical representation of SAFER as applied to cleaned up Data Set 2, after errors have been corrected. This method is used in extended self-modeling curve resolution, and resolves source compositions between the inner and outer boundaries of the 'feasible region'.

such Figure 12.10 shows the data projected onto a flat plane perpendicular to principal component 1. The SAFER method begins by defining the 'feasible region' where the simplex vertices and edges may reside. The inner boundary of the feasible region is defined by the convex hull of the data cloud (Figure 12.10). The non-negativity constraints on the analytes define the outer boundary of the 'feasible region'. Each of the lines through the gray shaded region in Figure 12.10 is the 'zero line' for a particular PCB congener, projected into three-PC space. Any potential source compositions that plots in the gray shaded region will have a negative composition for at least one analyte.

For a three-component system such as this, a feasible mixing model may be defined by direct inspection of the data plotted in principal component space, and manually located within the feasible region (this method is termed SAFER3D). A method of resolving higher dimensional mixing models has recently been described (Kim and Henry, 1999). That method calls on the use of additional explicit physical constraints. Examples of additional constraints may include (1) total mass of samples, (2) *a priori* knowledge of a subset of contributing sources, (3) upper and lower limits on ranges or ratios of analyte compositions, or (4) constraints based on laws of chemistry (Kim and Henry, 1999). As was the case for the unique vector iteration method, SAFER has been applied primarily in source apportionment studies in air (Henry *et al.*, 1994; Henry *et al.*, 1997). The original ESMCR software package SAFER has recently been revised, and will soon be available under the name UNMIX, available through Dr Ronald C. Henry, West Hills, California. The source compositions resolved by SAFER3D are shown in Figure 12.8. As for PVA and TTFA, the resolved source compositions are in good agreement with the true source profiles.

12.4 SUMMARY

Environmental forensics, by its very nature, involves analysis of complicated chemical data sets. These data typically contain information for many samples and with many measured chemicals. By definition, we are working in a multivariate, multidimensional world, and we must bring multivariate statistical methods to bear on these problems. However, the nature of environmental forensics investigations is such that we usually do not fully understand the systems under study. Seldom can we or should we apply classical statistical methods related to formal hypothesis testing. Rather, we must employ multivariate methods of exploratory data analysis. These methods must have special features: (1) they must be able to handle mixtures; (2) the results must be interpretable in a scientific context; (3) they should minimize *a priori* assumptions of data distributions and source chemistry; and (4) they must be able to handle systems of more than three sources. These conditions greatly limit our choices. The methods presented in this chapter (PCA and self-training receptor modeling methods) satisfy these criteria.

All of the methods discussed can produce spurious results when faced with bad data. A single highly aberrant measurement can significantly disrupt the variance structure of the data. This is the Achilles heel of multivariate procedures that depend on variance. Thus one of the most crucial steps in the data analysis process is vigilant outlier detection and data cleaning. PCA and receptor modeling must include systematic and objective procedures for evaluation of the quality of the data, and this is often best accomplished through the use of goodness-of-fit scatter plots. Once the data are cleaned, determination of the number of sources is relatively straightforward. If this step is effectively done, then the various methods used to determine the number of sources are usually consistent.

A hierarchy of procedures can then be used to analyze the cleaned data. PCA, the earliest of the procedures discussed, works best in simple cases, where there are few sources contributing to the system, and there is limited mixing between sources. If an initial PCA indicates the presence of mixtures, it usually best to move to a data analysis method capable of resolving the nature of that mixture. Three methods have been presented here: PVA, TTFA, and SAFER. Each of these methods is effective. TTFA's unique vector iteration method is the least attractive of these, because it still requires the user to evaluate a large number of candidate source profiles for feasibility.

Finally, in our eagerness to apply such methods to environmental chemical data, and our striving to develop more quantitative and rigorous methods, we must not lose sight of the chemical and scientific context of our project. Every time we find a surprise in a data set (which will be quite often, if we are doing our job correctly) we must relate that information back to the full context of study. For example, if an outlier is indicated on a scatter plot, we must go back to the chemist, or to the data entry technician, or to the field-team leader, and ask what might have caused that sample to be unique. The appropriate data analysis decision (e.g., deletion, modification, collection of new samples) will vary, depending on what we learn in those discussions. The data analysis may bring it to our attention, but questions of cause, source and scientific significance can rarely be answered in the context of numerical data analysis. More often, we are better served to consider such questions in the context of industrial history, chemistry, geology, and biology. We may borrow methods from mathematics and statistics, but we must remain principally environmental scientists.

ACKNOWLEDGMENTS

Thank you to Ron Henry for providing the software to run SAFER. Thank you to Chris Evans, Barry Lavine and Jason Ritter for critically reviewing the PVA formalism section. We also express our gratitude to Robert Wagner and Jeff Chiarenzelli for discussions regarding plausible PCB coelution and data error scenarios.

REFERENCES

Bedard, D.L. and Quensen, J.F. (1995) Microbial reductive dechlorination of polychlorinated biphenyls. In *Microbial Transformation and Degradation of Toxic Organic Chemicals* (Young, L.Y. and Cerniglia, C.E., eds), pp. 127–216. Wiley-Liss, New York.

Bezdek, J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Bezdek, J.C., Ehrlich, R., and Full, W.E. (1984) FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences* 10, 191–203.

Borbély-Kiss, I., Koltay, E., and Szabo, G. (1993) Apportionment of atmospheric aerosols collected over Hungary to sources by target transformation factor-analysis. *Nuclear Instruments and Methods in Physics Research* B75, 287–291.

Bright D.A., Cretney, W.J., MacDonald, R.W., Ikonomou, M.G., and Grundy, S.L. (1999) Differentiation of polychlorinated dibenzo-p-dioxin and dibenzofuran sources in coastal British-Columbia, Canada, *Environmental Toxicology and Chemistry* 18, 1097–1108.

Cattell, R.B. (1966) The scree test for the number of factors. *Multivariate Behavioral Research* 1, 245–276.

Chang, S.N., Hopke, P.K, Gordon, G.E., and Rheingrover, S.W. (1988) Target transformation factor analysis of airborne particulate samples selected by wind-trajectory analysis. *Aerosol Science and Technology* 8, 63–80.

Chiarenezelli, J.R., Scrudato, R.J., and Wunderlich, M.L. (1997) Volatile loss of PCB Aroclors from subaqueous sand. *Environmental Science and Technology* 31, 587–602.

Davis, J.C. (1986) Statistics and Data Analysis in Geology. John Wiley & Sons, New York.

Deane, J.M. (1992) Data reduction using principal components analysis. In *Multivariate Pattern Recognition in Chemometrics* (Brereton, R.G., ed.), pp. 125–177 Elsevier, New York.

Doré, T.J., Bailey, A.M., McCoy, J.W., and Johnson, G.W. (1996) An examination of organic/carbonate-bound metals in bottom sediments of Bayou Trepagnier, Louisiana. *Transactions of the Gulf Coast Association Geology Society* 46, 109–116.

Draper, N.R. and Smith, H. (1981) Applied Regression Analysis. John Wiley & Sons, New York. 709 p.

Eastman, H.T. and Krzanowski, W.J. (1982) Cross-validatory choice of the number of components from a principal components analysis. *Technometrics* 24, 73–77.

Ehrlich, R. and Full W.E. (1987) Sorting out geology, unmixing mixtures. In *Use and Abuse of Statistical Methods in Earth Sciences* (Size, W., ed.), pp. 34–46. Oxford University Press.

Ehrlich, R., Wenning, R.J., Johnson, G.W., Su, S.H., and Paustenbach, D.J. (1994) A mixing model for polychlorinated dibenzo-*p*-dioxins and dibenzofurans in surface sediments from Newark Bay, New Jersey using polytopic vector analysis. *Archives of Environmental Contamination and Toxicology* 27, 486–500.

Exner, O. (1966) Additive physical properties. Collection of Czechoslovak Chemical Communications 31**,** 3222–3253.

Frame, G.M. (1999) Improved procedure for single DB-XLB column GC-MS-SIM quantitation of PCB congener distributions and characterization of two different preparations sold as ‘Aroclor 1254’. *Journal of High Resolution Chromatography* 22, 533–540.

Frame, G.M., Cochran, J.W., and Bøwadt, S.S. (1996) Complete PCB congener distributions for 17 Aroclor mixtures determined by 3 HRGC systems optimized for comprehensive, quantitative, congener-specific analysis. *Journal of High Resolution Chromatography* 19, 657–668.

Full, W.E., Ehrlich, R., and Klovan, J.E. (1981) Extended Qmodel – objective definition of external end members in the analysis of mixtures. *Journal of Mathematical Geology* 13, 331–344.

Full, W.E., Ehrlich, R., and Bezdek, J.C. (1982) Fuzzy QModel – A new approach for linear unmixing. *Journal of Mathematical Geology* 14, 259–270.

Gemperline, P.J. (1984) A priori estimates of the elution profiles of pure components in overlapped liquid chromatography peaks using target transformation factor analysis. *Journal of Chemical Information and Computer Sciences* 24, 206–212.

Gordon, G.F. (1988) Receptor models. *Environmental Science and Technology* 22, 1132–1142.

Grung, B. and Kvalheim, O.M. (1994) Rank determination of spectroscopic profiles by means of cross validation: the effect of replicate measurements on the effective degrees of freedom. *Chemometrics and Intelligent Laboratory Systems* 22, 115–125.

Harman, H.F. (1960) *Modern Factor Analysis*. University Chicago Press.

Henry, R.C. and Kim, B.M. (1990) Extension of self-modeling curve resolution to mixtures of more than three components. Part 1: Finding the basic feasible region. *Chemometrics and Intelligent Laboratory Systems* 8, 205–216.

Henry, R.C., Lewis, C.W., and Collins, J.F. (1994) Vehicle related hydrocarbon source compositions from ambient data: the GRACE/SAFER method. *Environmental Science and Technology* 28, 823–832.

Henry, R.C., Spiegelman, C.H., Collins, J.F., and Park, J.F. (1997) Reported emissions of organic gases are not consistent with observations. *Proceedings of the National Academy of Sciences*, USA 94, 6596–6599.

Henry, R.C., Park, E.S., and Spiegelman, C.H. (1999) Comparing a new algorithm with the classic methods for estimating the number of factors. *Chemometrics and Intelligent Laboratory Systems* 48, 91–97.

Hopke, P.K. (1989) Target transformation factor analysis. *Chemometrics and Intelligent Laboratory Systems* 6, 7–19.

Hopke, P.K. (1991) An introduction to receptor modeling. *Chemometrics and Intelligent Laboratory Systems* 10, 21–43.

Imbrie, J. (1963) *Factor and Vector Analysis Programs for Analyzing Geologic Data*. Office of Naval Research. Tech Report No. 6. 83 pp.

Jarman, W.M., Johnson, G.W., Bacon, C.E., Davis, J.A., Risebrough, R.W., and Ramer, R. (1997) Levels and patterns of polychlorinated biphenyls in water collected from the San Francisco Bay and Estuary, 1993–1995. *Frenius' Journal of Analytical Chemistry* 359, 254–260.

Johnson, G.W. (1997) Application of Polytopic Vector Analysis to Environmental Geochemistry Problems. PhD Dissertation. University of South Carolina. Columbia, SC.

Johnson, G.W., Jarman, W.J., Bacon, C.E., Davis, J.A., Ehrlich, R., and Risebrough, R.W. (2000) Resolving polychlorinated biphenyl source fingerprints in suspended particulate matter of San Francisco Bay. *Environmental Science and Technology* 34, 552–559.

Kaiser, H.F. (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 187–200.

Kim, B.M. and Henry, R.C. (1999) Extension of self-modeling curve resolution to mixtures of more than 3 components Part 2 – finding the complete solution. *Chemometrics and Intelligent Laboratory Systems* 49, 67–77.

Klovan, J.E. (1968) Q-mode factor analysis program in FORTRAN-IV for small computers. Kansas Geological Survey Computer Contribution 20, pp. 39–51.

Klovan, J.E. and Imbrie, J. (1971) An algorithm and fortran-IV program for large-scale Q-mode factor analysis and calculation of factor scores. *Journal of Mathematical Geology* 3, 61–77.

Klovan J.E. and Miesch, A.T. (1976) EXTENDED CABFAC and QMODEL, computer programs for Q-mode factor analysis of compositional data. *Computers and Geosciences* 1, 161–178.

Malinowski, E.R. (1977) Determination of the number of factors and the experimental error in a data matrix. *Analytical Chemistry* 49, 612–617.

Malinowski, E.R. (1991) *Factor Analysis in Chemistry*. John Wiley & Sons, New York.

Manson, V. and Imbrie, J. (1964) FORTRAN program for factor and vector analysis of geological data using an IBM 7090 or 7094 computer system. Kansas Geological Survey Special Distribution Publication 13.

Miesch, A.T. (1976a) Q-mode factor analysis of geochemical and petrologic data matrices with constant row sums. *Geological Survey Prof. Paper*, 574-g, pp. 1–47.

Miesch, A.T. (1976b) Interactive computer programs for petrologic modeling with Extended Q-mode factor analysis. *Computers and Geosciences* 2, 439–492.

Moro, G., Lasagni, M., Rigamonti, N., Cosentino, U., and Pitea, D. (1997) Critical review of the receptor model based on target transformation factor analysis. *Chemosphere* 35, 1847–1865.

Ozeki, T., Koide, K., and Kimoto, T. (1995) Evaluation of sources of acidity in rainwater using a constrained oblique rotational factor analysis. *Environmental Science and Technology* 29, 1638–1645.

Reyment, R.A. and Jöreskog, K.G. (1993) *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, Cambridge.

Roscoe, B.A. and Hopke, P.K. (1981) Comparison of weighted and unweighted target transformation rotations in factor analysis. *Computers and Chemistry* 5, 1–7.

Tanabe, S., Kannan, N., Subramanian, A., Watanabe, S., and Tatsukawa, R. (1987) Highly toxic coplanar PCBs: Occurrence, source, persistence and toxic implication to wildlife and humans. *Environmental Pollution* 47, 147–163.

Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal components analysis models. *Technometrics* 20, 397.

Xie, Y.L., Hopke, P.K., and Paatero, P. (1998) Positive matrix factorization applied to a curve resolution problem. *Journal of Chemometrics* 12, 357–364.

Zhou, D., Chang, T., and Davis, J.C. (1983) Dual extraction of *R*-mode and *Q*-mode factor solutions. *Journal of Mathematical Geology* 15, 581–606.